Universidade Nova de Lisboa

Faculdade de Ciências e Tecnologia

Departamento de Informática

Data Warehouse para Dados Ambientais:

Estudo de Requisitos e Proposta de Modelação Multi-Dimensional.

Ana Sofia Carapinha da Cunha Lopes

asclopes@gmail.com

Dissertação apresentada na Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa para a obtenção do Grau de Mestre em Engenharia Informática.

Orientador:

Professor Doutor João Moura Pires

Lisboa, 30 de Outubro de 2006

Preâmbulo

Esta dissertação foi orientada pelo Prof. Doutor João Moura Pires e tem como objectivo a obtenção do grau de mestre em Engenharia Informática, curso administrado pelo Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa[1].

Esta dissertação tem como tema a problemática da modelação multi-dimensional, aplicada ao caso específico dos dados ambientais que, pela sua variedade, abrangência e amplitude de utilizações, representam um desafio muito interessante.

Esta tese foi enquadrada num protocolo entre o Instituto do Ambiente[2] e o Departamento de Informática da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, tendo sido o IA o principal fornecedor dos dados e do conhecimento adquirido ao longo desta tese, com base nos quais foi efectuado o trabalho aqui apresentado.

Nesta tese de mestrado serão focados principalmente os dados relativos aos agentes do ambiente. A informação associada aos processos IA, na sua componente de gestão, não será incluída no universo de estudo.

[1] http://www.di.fct.unl.pt
[2] http://www.iambiente.pt

Título da Dissertação:

Data Warehouse para Dados Ambientais: Estudo de Requisitos e Proposta de Modelação Multi-Dimensional.

Autor:

Ana Sofia Carapinha da Cunha Lopes

Palavras Chave:

Data Warehouse

Sistemas de Apoio à Decisão (SAD)

Metadados

Modelação conceptual

Qualidade dos dados

Modelação multi-dimensional

Data Warehouse para dados ambientais

Keywords:

Data Warehouse

Decision Support Systems (DSS)

Metadata

Conceptual modeling

Data quality

Multi-dimensional modeling

Environmental Data Warehouse

Agradecimentos

Ao Bruno, tenho a agradecer todas as vezes em que se encarregou das tarefas triviais que eu não conseguia fazer, e mesmo pela ajuda que me dava sempre que podia. Tendo ao meu lado uma pessoa assim, claramente só me apetece sorrir o dia inteiro.

Aos meus pais e irmãs, que me possibilitaram chegar à fase de realização de uma dissertação de mestrado, não tenho palavras suficientes para agradecer.

Aos meus amigos Susana e João, companheiros de luta durante todo o mestrado, muito obrigado pela empatia sentida que demonstraram, e pelo apoio que me deram. Que a Vida lhes sorria sempre, cheia de Felicidade.

Ao Jony, pela paciência e tempo que dispendeu no apoio à realização desta tese, bem como pela energia que demonstrou, só tenho agradecimentos a dar.

E finalmente, à Isa, sempre cheia de vigor, muito obrigada pela compreensão, suporte e paciência, principalmente nas alturas complicadas.

Sumário

Durante os últimos anos tem aumentado a atenção dedicada à análise e utilização da informação ambiental, dado que as estratégias políticas para a preservação do ambiente, gestão de recursos e prevenção de desastres ambientais que têm vindo a ser adoptadas assim o exigem. Surge então a necessidade de integrar e transformar os dados ambientais recolhidos por diversas entidades, de forma a que seja possível obter informação útil para permitir desenvolver e implementar as estratégias ambientais.

Neste sentido, o Instituto do Ambiente (IA), um organismo do Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional (MAOTDR), entidade encarregue da coordenação e planeamento na área de gestão do ambiente em Portugal, sentiu a necessidade de implementar um Data Warehouse que lhe permitisse integrar a informação relativa à actividade humana, a qual tem a responsabilidade de recolher e consolidar.

Esta informação inclui várias componentes ambientais (ar, água, solo, resíduos) provenientes de diferentes origens. Porque a recolha destes dados está em fase de desenvolvimento e melhoria, para fazer face às novas necessidades de análise e obtenção da informação, é imprescindível que o modelo multi-dimensional seja abrangente, e acomode facilmente as eventuais alterações que forem sendo efectuadas, numa perspectiva de extensibilidade.

Assim, esta dissertação pretende propor um modelo multi-dimensional genérico e flexível, sendo apresentadas várias alternativas de implementação e a justificação das opções tomadas. É ainda efectuada uma análise de robustez para este modelo, que permitirá avaliar qual a sua capacidade de abrangência e expansão, e é apresentado um pequeno protótipo (modelo e exploração) cujo âmbito teve em conta os dados já disponíveis no IA, e as ferramentas utilizadas nesta entidade, numa perspectiva de máximo reaproveitamento dos recursos.

Abstract

Over the last few years the focus in analysis and use of environmental information has increased, due to the adoption of politic strategies for environment preservation, resource management and environmental catastrophes prevention. This evolved into the need of integrating and transforming environmental data collected by several entities, allowing to obtain usefull information to develop and implement environmental strategies.

In this direction, the Institute for Environment (IA in the original designation), an organism from Environment, Territory Order and Regional Development Ministery (MAOTDR in the original designation), responsable for the coordination and planning on environment area in Portugal, felt the need to implement a *Data Warehouse*, to allow integration of information about human activity, whose collection and consolidation is IA's responsability.

This information includes several environmental components (air, water, soil, residues), from diferent sources. These data collections are constantly improving, to face new needs of analisys and reporting. So, the multidimensional model has to be flexible to accomodate new changes, in a extensible way.

This document's goal is to present a generic and flexible multidimensional model, with several implementation alternatives, and justify the chosen options. It is also presented the robustness analisys of the model, which will allow to evaluate it's enclosure and extensibility. A small prototype is presented (model and deployment), whose scope was the available data and tools of IA, in order to maximize internal resource exploitation.

Acrónimos

DW Data Warehouse

SAD Sistemas de Apoio à Decisão

ETL Extracting, Transforming and Loading

SIA Sistemas de Informação Ambiental

MAOTDR Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional

IA Instituto do Ambiente

EPER European Pollutant Emission Register

PCIP Prevenção e Controlo Integrado da Poluição

COV Compostos Orgânicos Voláteis

SCD Slowly Changing Dimensions

Conteúdo

Acrónimos **x**

1 Introdução **1**

1.1 Enquadramento 5

1.2 Objectivos e Contributos 7

1.3 Organização da dissertação 8

2 Sistemas de Apoio à Decisão e *Data Warehouse* **9**

2.1 Sistemas de Apoio à Decisão 10

2.2 Modelos de referência *Data Warehouse* 11

2.3 Modelação multidimensional 12

2.3.1 Dimensão tempo 15

2.3.2 Dimensões lentamente alteráveis 16

2.4 Processos de Extracção, Transformação e Carregamento (ETL) 19

2.4.1 Qualidade dos Dados 20

2.4.2 Rastreabilidade dos dados 23

2.5 Metadados 25

2.5.1 Common Warehouse Metamodel (CWM) 27

2.6 Modelação Conceptual - YAM^2 32

3 Trabalhos Relacionados - Sistemas de Informação Ambiental **39**

3.1 Indicadores e relatórios de estado do ambiente 42

3.1.1 Indicadores do ambiente 42

3.1.2 Relatório do estado do ambiente 43

3.1.3 Recolha de informação sobre o Estado do Ambiente 45

3.2 Sistemas de Apoio à Decisão Ambiental (SADA) 46

3.2.1 Sistemas de Apoio à Decisão Espaciais(SADE) 47

3.3 Casos de estudo . . . 49
3.3.1 Le Select . . . 50
3.3.2 Projecto SIMAGE . . . 53
3.3.3 Instalação Pantex . . . 56
3.3.4 Envirofacts . . . 58
3.4 Nota finais . . . 60

4 Modelo conceptual **61**
4.1 Conceitos do domínio a modelizar . . . 62
4.2 Âmbito . . . 66
4.3 Factores de modelação . . . 68
4.4 Sub-modelos considerados . . . 69
4.4.1 Dados das Emissões de Poluentes . . . 71
4.4.2 Dados de Produção . . . 82
4.4.3 Dados de Funcionamento . . . 84
4.4.4 Dados de COV . . . 86
4.4.5 Dados de descarga de águas residuais . . . 88
4.4.6 Dados de FontesPontuais . . . 92
4.4.7 Dados Socio-Demográficos . . . 96
4.5 Conclusões . . . 98

5 Protótipo **99**
5.1 Âmbito do protótipo . . . 100
5.2 Descrição do modelo físico do protótipo . . . 102
5.2.1 Emissões de Poluentes . . . 102
5.2.2 Dados de produção . . . 106
5.3 Processo de implementação . . . 107
5.4 Resultados obtidos . . . 109
5.5 Conclusões . . . 113

6 Conclusões e trabalho futuro **115**
6.1 Resumo . . . 116
6.2 Conclusões . . . 116
6.3 Trabalho futuro . . . 118

7 Anexos **121**
7.1 Conclusões do relatório de análise dos limiares . . . 122
7.2 Matriz de composição dos relatórios criados . . . 125

Lista de Figuras

1.1 Arquitectura do sistema de informação IA 5

2.1 Abordagens *top-down* e *bottom-up* 12
2.2 Exemplo de um esquema em estrela 13
2.3 Exemplo de um esquema *snowflake* 14
2.4 Metamodelo CWM: estrutura de pacotes 27
2.5 Metamodelo multidimensional CWM: classes e associações 29
2.6 Metamodelo OLAP CWM: classes e associações 30
2.7 Apresentação de um grupo de ferramentas de Data Warehouse com um repositório de informação partilhado (à esquerda), e do objectivo pretendido com o CWM, possibilitando troca de informação entre elas (à direita) 31
2.8 YAM: Estrutura dos nós YAM^2 34
2.9 YAM^2: Exemplo de representação da dimensão "Cliente" 35
2.10 YAM^2: Exemplo de representação das células num facto com as dimensões "Cliente" e "Produto" 35
2.11 YAM: Representação das relações suportadas 36

3.1 Diagrama do ciclo do carbono [SS] 40
3.2 Interacção dos componentes que se relacionam com o Ambiente 41
3.3 Arquitectura geral do Le Select 52
3.4 Esquema conceptual do protótipo SIMAGE 55
3.5 Arquitectura dos componentes do DWA da instalação Pantex 57

4.1 Componentes da actividade industrial 62
4.2 Grafo de conceitos associados à instalação 65
4.3 Componentes da actividade industrial 69

4.4 Diagrama YAM^2 de nível conceptual mais elevado para a estrela das emissões detalhadas . . . 71
4.5 Diagrama YAM^2 de nível conceptual mais elevado para a estrela das emissões agregadas . . . 73
4.6 Estrutura da dimensão Instalação . . . 73
4.7 Exemplo de uma implementação em *bridge* para as CAE secundárias da instalação 75
4.8 Diagrama YAM^2 de nível conceptual intermédio para os dados das emissões detalhadas . . . 76
4.9 Diagrama YAM^2 de nível conceptual intermédio para os dados das emissões agregadas . . . 76
4.10 Diagrama YAM^2 de nível conceptual superior para os dados de produção . . . 82
4.11 Diagrama YAM^2 de nível conceptual intermédio para os dados de produção . . . 83
4.12 Diagrama YAM^2 de nível conceptual superior para os dados de funcionamento . 84
4.13 Diagrama YAM^2 de nível conceptual intermédio para os dados de funcionamento 85
4.14 Diagrama YAM^2 de nível conceptual superior para os dados de COV . . . 87
4.15 Diagrama YAM^2 de nível conceptual intermédio para os dados de COV . . . 88
4.16 Diagrama YAM^2 de nível conceptual superior para os dados de descargas de águas residuais . . . 89
4.17 Diagrama YAM^2 de nível conceptual intermédio para os dados de descargas de águas residuais . . . 91
4.18 Diagrama YAM^2 de nível conceptual mais elevado para os dados relativos a equipamentos contribuintes . . . 93
4.19 Diagrama YAM^2 de nível conceptual intermédio para os dados relativos a equipamentos contribuintes . . . 94
4.20 Diagrama YAM^2 de nível conceptual mais elevado para os dados relativos a fontes pontuais . . . 95
4.21 Diagrama YAM^2 de nível conceptual intermédio para os dados relativos a fontes pontuais . . . 95
4.22 Diagrama YAM^2 de nível conceptual superior para os dados demográficos . . . 97
4.23 Diagrama YAM^2 de nível conceptual intermédio para os dados demográficos . . 97

5.1 Estrutura das estrelas que compõem o protótipo, previstas no modelo conceptual 101
5.2 Esquema físico das emissões detalhadas . . . 103
5.3 Esquema físico das emissões agregadas . . . 105
5.4 Esquema físico para implementação das análises quanto à conformidade da lista de poluentes . . . 105

5.5 Esquema físico dos dados de produção 107
5.6 Análise do número de instalações que relatam cada poluente 110
5.7 Evolução 2002-2004 dos dados de emissões relatados 112
5.8 Emissões relatadas e emitidas para o CO2, em 2004 112

7.1 Matriz de relatórios 125

Lista de Tabelas

4.1 Descritores necessários para a compreensão das medidas dos factos 78
4.2 Medidas do facto **Dados_Em** 79
4.3 Medidas do facto **Dados_EmAgreg** 81
4.4 Descrição do facto **Dados_Prod** 83
4.5 Medidas do facto **Dados_Func** 86
4.6 Medidas do facto **Dados_COV** 88
4.7 Medidas do facto **Desc_AgResid** 92
4.8 Medida do facto **Dados_EqContrib** 94
4.9 Medidas do facto **Dados_FontPont** 96
4.10 Medidas do facto **Dados_SocioDemog** 98

7.1 Poluentes para os quais poderá ser tomada em consideração a diminuição dos valores dos limiares 123

1

Introdução

Neste capítulo será apresentado o contexto em que se insere a presente dissertação e será feita a descrição do universo e das temáticas abrangidas no âmbito deste trabalho.

O conceito de **Desenvolvimento Sustentável** é, normalmente, definido como o desenvolvimento que procura satisfazer as necessidades da geração actual sem comprometer a capacidade das gerações futuras de satisfazerem as suas próprias necessidades.

Isto significa que tem de se possibilitar que as pessoas, agora e no futuro, atinjam um nível satisfatório de desenvolvimento social e económico e de realização humana e cultural, fazendo ao mesmo tempo um uso razoável dos recursos da Terra e preservando as espécies e os habitats naturais [POR].

O desenvolvimento sustentável assenta em 3 eixos: ambiental, social, e económico, só podendo ser alcançado se estes três eixos evoluírem de forma harmoniosa.

Para que se possa avaliar o impacto da evolução de cada um destes eixos nos outros, é necessário haver análise de informação, de forma a serem tomadas decisões ponderadas e sustentadas.

Existem duas vertentes principais a serem consideradas no âmbito da análse de dados ambientais, o eixo onde se enquadra esta dissertação:

- impacto ambiental da actividade humana - tem como principal objectivo analisar directamente os impactos da actividade humana no ambiente, através da medição, estimativa e cálculo das emissões efectuadas para o meio.
- estado do ambiente - pretende avaliar e prever os níveis de poluição e qualidade do ar, água e solos através da recolha e tratamento de dados provenientes de redes de monitorização ou de relatórios técnicos e estudos efectuados por diversas organizações.

Para a realização destas análises, é necessário que existam disponíveis dados processáveis e informação relevante, motivo pelo qual a interacção de dados e informação nos sistemas de informação ambiental se tem tornado um foco de atenção durante os últimos anos.

As próprias estratégias políticas para a preservação, gestão dos recursos naturais e a prevenção de desastres adoptadas por vários países exigem cada vez mais a disponibilidade de informação, pois necessitam de uma análise prévia antes da sua implementação, utilizando dados ambientais recolhidos a partir de entidades públicas e centros de investigação para validar e verificar o seu provável impacto. Normalmente estes dados são disponibilizados através de diferentes recursos, e portanto necessitam de ser integrados e transformados em informação que pode ser utilizada.

No contexto do território português o Instituto do Ambiente (IA), um organismo do Ministério do Ambiente, do Ordenamento do Território e do Desenvolvimento Regional (MAOTDR) assume um papel de coordenação geral, de harmonização de procedimentos e de utilização coerente dos instrumentos normativos requeridos para aplicação das políticas ambientais estabelecidas a nível regional ou local, criando, sempre que tal se justifique, parcerias com os organismos de coordenação regional ou autárquica, ou directamente com as autarquias,

para a realização dos seus objectivos.

Entre outras atribuições, o IA é responsável por [dR03]:

- apoiar a definição da política ambiental e acompanhar a execução e avaliação dos resultados alcançados;
- assegurar, em sede de licenciamento ambiental, a adopção das medidas de prevenção e controlo integrado da poluição pelas instalações por elas abrangidas;
- promover, coordenar e apoiar a concretização de estratégias, planos e programas nacionais de desenvolvimento sustentável e as que se referem a matérias de natureza global, nomeadamente as que respeitam às alterações climáticas, à protecção da camada de ozono, à limitação das emissões nacionais de poluentes atmosféricos, à avaliação de impactos num contexto transfronteiriço e à segurança biológica;
- assegurar a recolha, tratamento e análise da informação relativa ao ambiente e elaborar o relatório do estado do ambiente;
- promover o acesso do público à informação ambiental e a participação do público na formulação e debate de políticas ambientais.

Para a realização destas funções, o IA tem um conjunto de sectores técnicos, dos quais se referem alguns.

- Avaliação de Impactos Ambientais (AIA) - é um instrumento preventivo da política de ambiente e do ordenamento do território que permite assegurar que as prováveis consequências sobre o ambiente de um determinado projecto de investimento sejam analisadas e tomadas em consideração no seu processo de aprovação.
- Prevenção e Controlo Integrados da Poluição (PCIP) - licenciamento ambiental obrigatório para as instalação consideradas como abrangidas pelo PCIP ou seja, aquelas onde se desenvolve uma ou mais das actividades consideradas no anexo I do Decreto-Lei 194/2000, de 21 de Agosto - ainda que essas actividades não constituam a principal actividade da instalação - e que cumpram certos requisitos quanto aos seus valores de capacidades de produção.

  Ao abrigo da implementação do European Pollutant Emission Register (EPER) em Portugal, que pretende estabelecer um registo europeu de dados comparáveis relativos às emissões de poluentes provenientes dos vários países, as instalações abrangidas pelo diploma PCIP deveriam relatar dados relativos às suas emissões.
- Compostos Orgânicos Voláteis (COV) - visa regular as emissões de COV para o ambiente, resultantes da aplicação de solventes orgânicos em certas actividades e instalações.
- Seveso [1] - em 1982 foi adoptada a directiva europeia, normalmente denominada "Direc-

[1] O nome "Seveso" tem origem na denominação de uma cidade italiana onde em 1976 se romperam tanques de

tiva Seveso", mais tarde substituída pela Directiva Seveso II, que por sua vez foi posteriormente extendida. A Directiva Seveso II aplica-se aos estabelecimentos industriais onde existam substâncias perigosas em quantidades superiores a determinados limites, e pretende controlar o perigo de grandes acidentes envolvendo estas substâncias.

- Comércio Europeu de Licenças de Emissão (CELE) - como forma de garantir o cumprimento eficaz dos seus objectivos relativos à redução de emissões de Gases com Efeito de Estufa (GEE), a União Europeia (UE) aprovou a directiva que cria este mecanismo. Este regime pretende ajudar os estados-Membros da UE a cumprirem os seus compromissos ao abrigo do Protocolo de Quioto ao menor custo possível, não implicando directamente novos objectivos ambientais.

Além destes sectores já referidos e que estão directamente relacionados com os processos ambientais, existe ainda um outro sector do IA que deve ser relevado devido à sua importância estratégica, que é o sector de "Acesso à Informação e Participação do Cidadão". Este sector tem como função promover a divulgação generalizada da informação sobre o ambiente, assegurar a concepção e a realização de acções de sensibilização, educação e formação dos cidadãos no domínio do ambiente em cooperação com as organizações não governamentais, e promover o desenvolvimento da estratégia nacional de educação ambiental.

Além do IA existem ainda outras organizações governamentais, igualmente tuteladas pelo MAOTDR, directamente relacionadas com o ambiente, embora em âmbitos mais específicos, das quais mencionamos algumas a título exemplificativo.

- Instituto da Água (INAG) - é a Autoridade Nacional da Água e tem como objectivo executar as políticas de recursos hídricos a nível nacional.

- Inspecção-Geral do Ambiente e do Ordenamento do Território (IGAOT) - serviço central de inspecção do ministério cuja actuação visa garantir, por parte de entidades públicas e privadas, o cumprimento das normas jurídicas nas áreas do ambiente, ordenamento do território e conservação da natureza.

Desde 1976, com a Convenção de Barcelona [Proa], que a obtenção de informação ambiental se manifesta como uma preocupação. Na realidade, um dos focos que desde cedo teve impacto a nível internacional é a temática do Ambiente. E, a forma de prestação de dados por parte dos vários países envolvidos para garantir uma base de informação útil e transversal tem sido amplamente discutida, como por exemplo nos relatórios publicados em 1999, 'A checklist for state of the environment reporting" [KAD99] e "A new model of environmental communication for Europe from consumption to use of information" [ddAICdT99]. Assim, uma das

uma indústria química, tendo sido libertados vários quilogramas da dioxina TCDD (2,3,7,8-tetraclorodibenzo-p-dioxina), uma substância venenosa e potencialmente cancerígena, na atmosfera e no solo. Esta ocorrência obrigou à evacuação de mais de 600 pessoas, e cerca de 2000 tiveram de receber tratamento para o envenenamento sofrido.

principais actividades do IA é exactamente o cumprimento das directivas europeias ao nível do fornecimento de dados ambientais, motivação suficiente para suportar o trabalho apresentado nesta dissertação.

Neste primeiro capítulo será apresentado o enquadramento onde se insere esta dissertação, os seus objectivos e contributos, e será explicada a estrutura deste documento.

1.1 Enquadramento

O IA tem efectuado ao longo dos últimos anos uma grande aposta na áreas das tecnologias de informação, tendo começado pela definição de uma arquitectura de Sistemas de Informação transversal e completa, e iniciado a informatização dos seus formulários de suporte à recolha de informação ambiental, conforme se apresenta na figura 1.1.

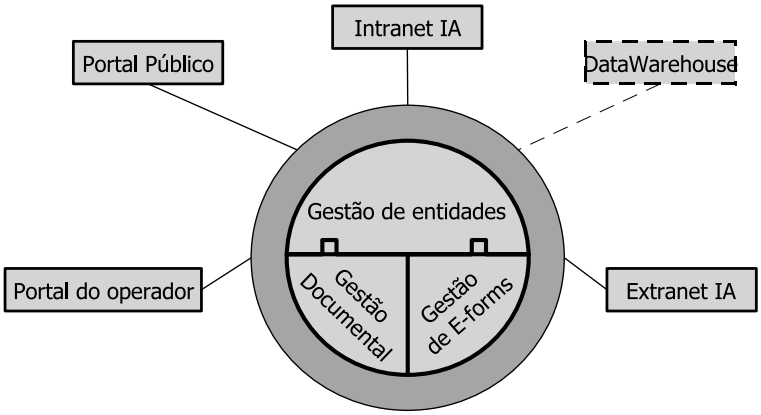

Figura 1.1: Arquitectura do sistema de informação IA

Esta arquitectura é constituída por vários componentes, que passamos a apresentar.

- O **portal do operador** é a interface de entrada para o registo das instalações industriais e para o fornecimento dos dados relativos às mesmas por parte dos seus operadores (entidades que exploram a referida instalação). Através deste interface são também iniciados grande parte dos processos ambientais a cargo do IA.
- O **portal público** exerce as funções de informação ao público, permitindo a difusão de publicações (como por exemplo os Relatórios de Estado do Ambiente), eventos, e mesmo compromissos legais, como prazos e legislação aplicável.
- Na **intranet IA** encontram-se as aplicações operacionais que suportam os vários processos (licenciamento, registos de dados, etc.).
- O componente *Data Warehouse* não foi ainda desenvolvido, mas como é uma das preocupações actuais do IA, foi já incluído no diagrama.
- A **extranet** do IA destina-se à comunicação com as entidades parceiras e outras organizações que constituem as autoridades competentes, como por exemplos as Comissões de

Coordenação e Desenvolvimento Regional(CCDR).

Toda a arquitectura se baseia na existência de um repositório central de entidades (**Gestão de Entidades**), onde são registados os dados de todas as instalações industriais e entidades com as quais o IA se relaciona.

Existe também um módulo para realizar a **gestão dos documentos** do IA e, finalmente, o módulo de **gestão dos** formulários electrónicos (**e-forms**) enviados pelos utilizadores.

Para a implementação de um Sistema de Apoio à Decisão (SAD) no IA poderiam ser considerados vários indicadores, genericamente agrupáveis nos seguintes grupos de informação:

- indicadores da actividade humana na perspectiva de impactos ambientais - corporizados por medições directas dos resultados da actividade humana;
- indicadores de estado do ambiente - são sobretudo baseados na recolha de informação de redes de monitorização ou estudos efectuados;
- indicadores de medição de eficiência dos processos da organização - estes processos corporizam a componente operacional do IA, e dividem-se em:
  - internos - processos de aquisição de equipamentos, recursos humanos, gestão das infraestruturas, etc;
  - externos - licenciamento ambiental das instalações industriais, avaliações do impacto ambiental, etc;
- indicadores de eficiência e impacto da função de educação e informação ao público - para o estudo destes indicadores teria de ser prevista a realização de sondagens ao público (inquéritos, estudos de opinião, etc.) que permitissem efectuar esta avaliação.

Na presente dissertação só será abrangido o primeiro ponto, ou seja, os indicadores da actividade humana na perspectiva de impactos ambientais. Note-se que os dados necessários para estes indicadores resultam muitas vezes dos processos externos ao IA, mas neste trabalho não pretendemos analisar a eficiência da implementação desses processos, mas sim o seu produto.

O trabalho apresentado nesta dissertação foi efectuado no âmbito de um protocolo entre o IA e a Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, cujo objectivo se centra na prestação de auxílio ao IA no início do desenvolvimento de um sistema de apoio à decisão. O objecto deste protocolo abrangeu várias vertentes, nomeadamente:

- apoiar o IA no desenvolvimentos das análises necessárias para a produção do *Overview Report*, a ser apresentado à União Europeia (UE) no âmbito do European Pollutant Emission Register (EPER);
- produzir um conjunto de requisitos analíticos, incluindo a especificação de indicadores, eixos de análise, relatórios e análises padrão, para serem aplicados aos dados recolhidos

no âmbito do EPER;

- verificação da conformidade dos limiares de poluentes definidos pela UE à realidade portuguesa, avaliando se as taxas de relato de dados à UE atingem os valores pretendidos (cerca de 90% das emissões produzidas), abrangendo o mínimo número de instalações possível. No anexo 7.1 apresenta-se o resumo das conclusões alcançadas.
- apresentação de uma proposta de modelação multi-dimensional a implementar no IA. Para além dos modelos Data Warehouse estarão incluídos os fluxos de *Extraction, Transformation* and *Loading* (ETL) bem como as regras de carregamento da informação;
- selecção das ferramentas analíticas a usar no IA, incluindo apoio na elaboração de cadernos de encargos e avaliação das propostas, caso venha a ser necessário;
- implementação dos modelos multi-dimensionais propostos e definição dos mecanismos de controlo de qualidade dos dados no Data Warehouse;
- apoio à implementação dos processos ETL;
- elaboração de um relatório final de avaliação do *Data Warehouse* entretanto implementado e a proposta de um *road-map*.

Na secção seguinte apresentam-se os objectivos considerados neste trabalho e os contributos relevantes que podem ser retirados do trabalho apresentado nesta dissertação.

1.2 Objectivos e Contributos

Tendo em conta os aspectos referidos nas secções anteriores, o estudo apresentado nesta tese tem de ser abrangente, para que a solução apresentada seja de fácil expansão, permita superar as dificuldades inerentes à diversidade e heterogeneidade da informação disponível e acomode facilmente novos requisitos, pois trata-se de um tema que começou recentemente a ser abordado no IA e consequentemente é inevitável a necessidade de actualizações.

Foram efectuadas várias pesquisas no sentido de tentar encontrar outros trabalhos efectuados na área de sistemas de apoio à decisão para dados do ambiente que pudessem ter interesse e valor acrescentado para a compreensão da temática em estudo. Como resultado dessas pesquisas foram encontrados os projectos apresentados na secção 3.3.

No entanto, não foi possível encontrar nenhum projecto cuja apresentação chegasse ao detalhe de discutir e analisar as opções de implementação dos modelos subjacentes, pelo que o trabalho apresentado nesta dissertação foi planeado e desenvolvido tendo como ponto de partida os pressupostos de expansibilidade e evolução já apresentados e o estudo da informação disponível em termos de dados do ambiente, que foi efectuado no próprio IA.

Tendo como base estas duas componentes foram definidos os requisitos de modelação que deveriam ser contemplados neste trabalho, apresentados na secção 4.3, e que serviram de base para a implementação de um protótipo no IA.

1.3 Organização da dissertação

Para a apresentação deste trabalho foram feitas algumas opções de escrita quanto à utilização de termos ténicos, tendo-se optado por não corrigir algumas expressões que são mais frequentemente conhecidas pelo seu original inglês, nomeadamente:

- Data Warehouse (DW);
- Star schema;
- Slowly Changing Dimensions (SCD);
- Data Marts.

Para a sigla ETL será utilizado o seu original inglês (Extraction, Transformation and Loading), pois considera-se mais facilmente reconhecível do que a utilização da versão portuguesa "Extracção, Transformação e Carregamento"

No capítulo onde nos encontramos presentemente foi efectuada a apresentação resumida do enquadramento, objectivos e contributos da dissertação. Estes pontos serão explicados com maior detalhe nos capítulos de desenvolvimento deste trabalho.

No capítulo 2 será feita referência à teoria relacionada com modelação multi-dimensional, metadados, processos ETL e modelação conceptual.

Aborda-se no capítulo 3 alguns projectos relacionados com a utilização de sistemas de apoio à decisão com dados do ambiente.

A seguir passa-se à exposição detalhada do trabalho desenvolvido no âmbito desta dissertação, sendo apresentados no capítulo 4 alguns conceitos do domínio e a descrição do modelo conceptual proposto e no capítulo 5 a descrição do protótipo implementado.

Por último serão apresentadas as conclusões e trabalho futuro.

2

Sistemas de Apoio à Decisão e *Data Warehouse*

Apresenta-se neste capítulo teoria relacionada com as temáticas de Sistemas de Apoio à Decisão e Data Warehouse.

Os Sistemas de Apoio à Decisão (SAD) emergiram da necessidade das organizações serem competitivas, eficientes e rentáveis. Para permanecerem competitivas, as empresas precisam de acelerar as suas decisões políticas, reagindo rapidamente à evolução do mercado, normalmente analisando e planeando políticas adequadas e estratégicas relacionadas com o seu negócio.

É, no entanto, reconhecido [HHD98] que não é possível antecipar todas as necessidades de informação, pois estas vão dependender da situação de negócio que estiver em causa.

Os responsáveis pela tomada de decisão precisam de rever dados sob diferentes pontos de vista e a diferentes níveis de detalhe antes de poderem decidir a forma de abordagem. Também precisam de detectar novas oportunidades de negócio e de analisar o desempenho da empresa, bem como antecipar eventuais situações que possam surgir de forma pró-activa, prevenindo em vez de simplesmente remediar.

Os SAD pretendem facilitar a utilização de dados, modelos e processos de decisão estruturada na actividade de tomada de decisão. Desde 1975 ([oISS96]) que se considera que os dados existentes numa organização são um recurso de informação importante e que deve ser gerido e explorado. No entanto, existem várias formas de se efectuar essa exploração.

2.1 Sistemas de Apoio à Decisão

Existem várias abordagens para apoio à decisão, como por exemplo, as àrvores de decisão [Ins], a tomada de decisão multi-critério (MCDM) [Cenb] e o processo analítico hierárquico (AHP) [Cena].

A própria utilização de técnicas de inteligência artificial, combinadas com gestão de modelos, é essencial para o apoio à decisão e é vista como uma das aproximações possíveis para os SAD mais avançados [HRR00].

O que estas abordagens têm em comum é que de uma forma geral se baseiam em modelos matemáticos, têm um uso restrito em termos da dimensão dos conjuntos de dados que suportam (apenas de pequena dimensão) e normalmente requerem a intervenção de peritos nas abordagens propostas para efectuar as configurações de base, para que as técnicas possam ser utilizadas em problemas concretos.

São também orientadas para responderem a determinadas questões, utilizando um determinado leque de dados, pois baseiam-se em modelos adaptados à resolução de problemas específicos. Portanto, quando surge a necessidade de responder a novas questões, utilizando novos tipos de dados, é necessário construir um novo modelo, adaptado ao novo problema.

Mas, cada vez mais a necessidade das organizações evoluiu no sentido de ser necessário tomar decisões com bases em grandes conjuntos de dados e, no contexto das grandes e médias organizações, haverem sistemas que permitissem a diferentes actores pouco familiariza-

dos com a utilização de computadores, tomarem decisões de forma rápida e simples.

Assim, no final da década de 70 [Pow03] surgiu uma nova abordagem, evoluída a partir dos sistemas de apoio à decisão baseados em modelos e que contribuiu para a melhoria das bases de dados relacionais, os Sistema de Informação Executiva (EIS) [Wikc]. Estes sistemas tinham como objectivo agrupar os dados da organização e providenciar os dados necessários aos responsáveis pela tomada decisão, como por exemplo volumes de vendas e estatísticas de pesquisas de mercado, sendo os primeiros sistemas de apoio à decisão baseados em dados (por contraposição aos sistemas baseados em modelos).

No início dos anos 90 os *Data Warehouse* (DW) e o *On-Line Analytical Processing* (OLAP) (que será apresentado na secção 2.3) começaram a alargar o âmbito dos EIS, e definiram uma categoria mais vasta de sistemas de apoio à decisão baseados em dados.

2.2 Modelos de referência *Data Warehouse*

Um DW tem os seguintes objectivos [KR02]:

- ser um repositório de informação seguro, pelo que tem de controlar o acesso à informação confidencial da organização;
- servir como base para a tomada de decisão, pelo que tem de ser abrangente e conter os dados necessários e suficientes para cumprir esta função.

Assim, um DW armazena dados transformados e integrados que podem abranger uma vasta área operacional da empresa e que são necessários para a tomada de decisão estratégica. Além disso, pode também conter dados de eventuais fontes externas que sejam relevantes para uma correcta e ponderada tomada de decisão.

Quando se discute a temática DW existem dois paradigmas essenciais a serem referidos, propostos por Bill Inmon [Sys] e Ralph Kimball [Ass].

A abordagem *top-down*, defendida por Inmon encara o DW como sendo parte do sistema de informação de negócio que abrange toda a organização [1ke]. Ou seja, o DW é à partida modelizado de forma a abranger toda a organização, contendo dados históricos detalhados, e a informação é armazenada de forma normalizada. Depois, são então criadas as estruturas de dados sectorizadas, necessárias para o suporte à decisão dentro de cada àrea de negócio, os *Data Marts*.

A abordagem *bottom-up* defendida por Kimball, considera que o DW é constituído pela aglomeração de todos os *Data Marts* dentro da organização, sendo que a informação está sempre armazenada num modelo dimensional. Ou seja, começa-se a desenvolver por partes e no fim integra-se tudo se for necessário.

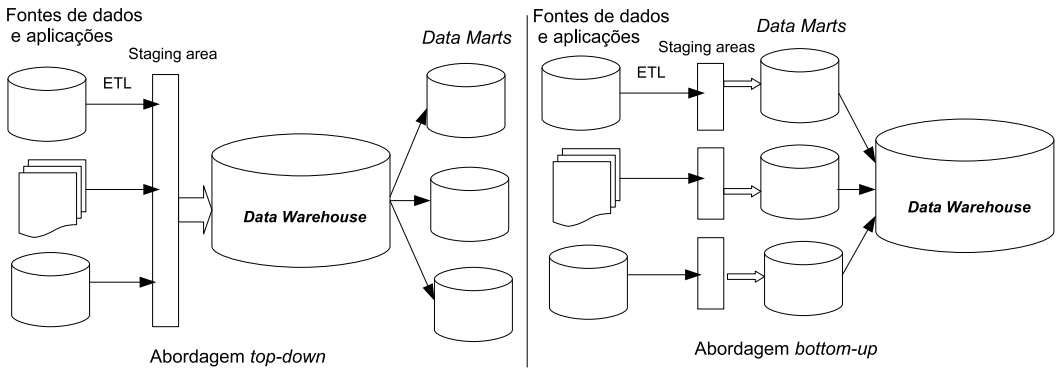

Figura 2.1: Abordagens *top-down* e *bottom-up*

Na figura 2.1 apresenta-se o diagrama das duas abordagens. Como se pode verificar, estas representam duas filosofias distintas, sendo que os DW de grande parte das organizações se aproximam mais da ideia de Kimball, porque na sua maioria começaram através de um esforço departamental, originando um Data Mart, e só depois evoluíram para um verdadeiro DW.

Qualquer que seja a abordagem escolhida, existem vários conceitos genéricos associados a um DW, que serão apresentados ao longo das secções seguintes, nomeadamente:

- modelação multidimensional - este nome deriva da noção de dimensão, ou agrupamentos de categorias. Nesta secção inclui-se a apresentação do OLAP e suas variantes (ROLAP, HOLAP e MOLAP);
- extracção, transformação e carregamento (ETL na sigla em inglês) de dados;
- metadados, ou informação sobre os dados.

2.3 Modelação multidimensional

Existem duas noções inerentes aos modelos multidimensionais: os factos ou métricas, os valores numéricos (contagens ou montantes que podem ser somados, usados para efectuar médias, maximizados e minimizados) que estão ser sujeitos a análise; e as dimensões, as perspectivas pelas quais se podem analisar os factos de forma vantajosa.

O objectivo do modelo multidimensional é caracterizar as actividades de negócio em termos de factos e dimensões, numa estrutura semelhante a um cubo, onde a célula do cubo representa o facto ou métrica orientada por cada uma das dimensões (faces) do cubo.

Genericamente os modelos multidimensionais são representados através de um *star schema*, ou esquema em estrela.

Um *star schema* é uma estrutura simples, com relativamente poucas tabelas e ligações bem definidas, que tem um tempo de resposta rápido para as pesquisas e é apresentado num esquema com o formato de uma estrela, facilmente compreendido mesmo por pessoas que não estejam familiarizadas com estruturas de bases de dados. O nome do esquema provém da

forma como é apresentado: no centro de um *star schema* estão as tabelas de factos contendo os valores numéricos, ou factos. Este esquema utiliza muitos dos componentes que se podem encontrar nos Diagramas de Entidades e Associações (DEA) tais como entidades, atributos, cardinalidades e chaves primárias.

Uma tabela de dimensão é constituída por uma única chave primária (que será referenciada pela tabela de factos), e uma ou mais colunas que contêm dados textuais sobre a dimensão. Temos como exemplos de dimensões a Geografia, Produto, Tempo. Na figura 2.2 apresenta-se um exemplo típico de um *star schema*, sendo o Produto, o Tempo, o Canal de Venda e o Cliente as suas dimensões, e Vendas a tabela de factos.

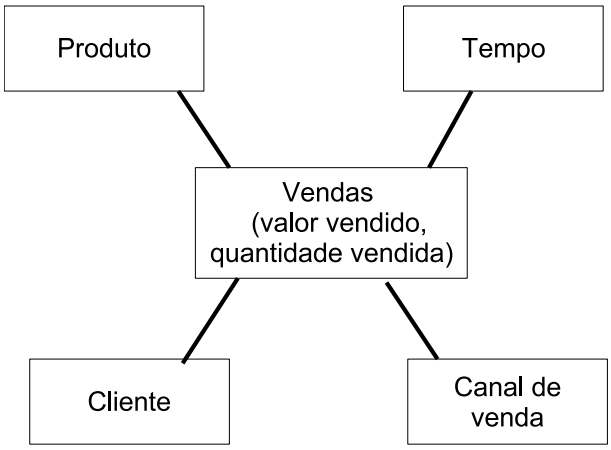

Figura 2.2: Exemplo de um esquema em estrela

Tipicamente as tabelas de factos têm grandes volumes de dados, enquanto as tabelas de dimensão tendem a ter um pequeno número de linhas. A vantagem principal desta aproximação é que o desempenho da junção de tabelas é melhorado quando uma tabela de grandes dimensões pode ser unida com várias mais pequenas. Frequentemente as tabelas de dimensões são suficientemente pequenas para serem completamente guardadas em memória.

Existem variantes de *star schema*, como por exemplo o *snowflake*, que adiciona estruturas hierárquicas às tabelas de dimensões. Na figura 2.3 apresenta-se o exemplo de um *snowflake*, com as mesmas dimensões do esquema anterior, mas onde as dimensões Cliente e Produto têm hierarquias, respectivamente para País e Fornecedor.

Sendo o DW o ponto central de integração e armazenagem de dados para suporte à decisão, pode-se dizer que o OLAP é a abordagem utilizada para se responder a questões analíticas que sejam dimensionais por natureza, com base na informação do DW.

Através do OLAP é possível efectuar operações de análise típicas de forma rápida e eficiente, como por exemplo a sumarização e agregação de grandes volumes de dados, filtragem, ordenação, comparação de vários conjuntos de dados, pesquisa de valores atípicos, descoberta de padrões e análises de tendência nos dados.

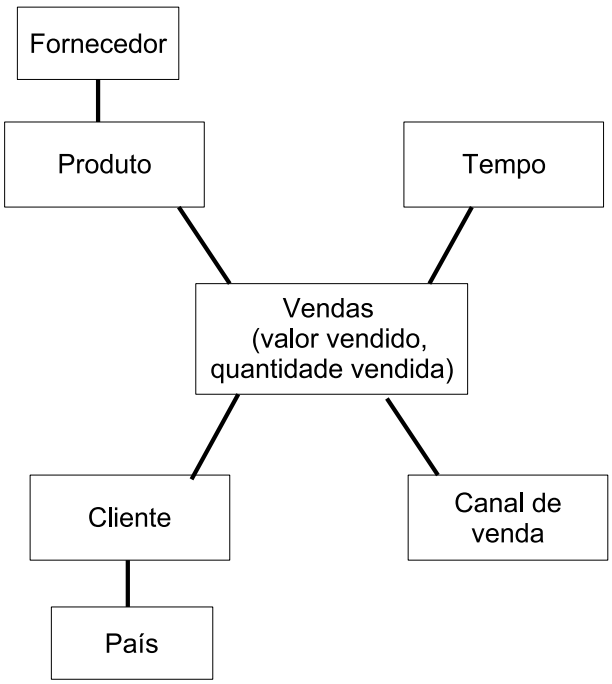

Figura 2.3: Exemplo de um esquema *snowflake*

A forma como os dados de uma base de dados multidimensional são fisicamente armazenados influencia muito o seu tempo de resposta. Existem várias estratégias de armazenamento de dados, cada uma adequada a uma situação diferente, designadas Multidimensional OLAP (MOLAP), Relational OLAP (ROLAP) e Hybrid OLAP(HOLAP).

O MOLAP utiliza uma estrutura multidimensional para armazenar as agregações e uma cópia dos dados base. Fornece um potencial tempo de resposta mais rápido, dependendo apenas do número e desenho das agregações do cubo. Em geral, é mais indicado para cubos que sejam usados frequentemente e que necessitem de tempos de resposta rápidos.

O ROLAP utiliza tabelas na base de dados relacional do DW para armazenar as agregações dos cubos. Em contraste com o armazenamento do tipo MOLAP, o ROLAP não armazena uma cópia dos dados base, acedendo à tabela de factos original quando necessário. O tempo de resposta do ROLAP é geralmente mais longo que o do MOLAP. É usado frequentemente com grandes volumes de dados que não precisem de ser frequentemente consultados (por exemplo, dados referentes a anos anteriores).

O HOLAP combina as características do MOLAP e do ROLAP. Os dados de agregação são armazenados em estruturas MOLAP e os dados de base restantes são deixados na base de dados relacional. Para pesquisas que acedam a dados agrupados, o HOLAP é equivalente ao MOLAP. Em pesquisas que usem a dados base com grande nível de detalhe, é necessário que estas vão buscar os dados à base de dados relacional, sendo portanto mais lentas que as anteriores. Os cubos armazenados como HOLAP são mais pequenos que os seus equivalentes MOLAP. São geralmente adequados para cubos que exijam rápidos tempos de resposta para

dados sumariados baseados numa grande quantidade de dados base.

2.3.1 Dimensão tempo

Os DW devem preservar a história. Isto implica que um DW deve preencher 3 requisitos [Kim03]:

- cada dado no DW tem de ter um período de validade determinado, com datas de início e fim bem definidas;
- se a descrição detalhada de uma entidade do DW mudou ao longo do tempo, cada versão da entidade tem de ser correctamente associada com as versões actuais das outras medidas e entidades no DW. Por exemplo, se um cliente efectuou uma compra há um ano atrás, a descrição que está associada à compra efectuada tem de estar em conformidade com os dados vigentes na altura;
- no entanto, tem de ser simultâneamente possível ver os dados ao longo do tempo, visualizando imagens instantâneas, através de relatórios periódicos, ou verificando o estado actual.

Por estes motivos, é comum adicionar-se estampilhas temporais a cada registo da tabela de factos. Mas, num esquema dimensional, normalmente adopta-se a utilização de uma dimensão tempo, contendo atributos classificativos (mês, trimestre, semestre), por vezes contendo hierarquias de classificação paralelas e que facilitam a navegação nos dados. Por exemplo, uma dimensão tempo pode conter simultaneamente a hierarquia relativa ao ano civil - Janeiro/Dezembro e a hierarquia sazonal - época alta, época baixa.

Nos sistemas multidimensionais é necessário garantir que, para qualquer data a que se referem os dados, seja possível associar uma valor da dimensão tempo aos mesmos. Podem mesmo existir registos com datas inválidas ou corrompidas e deve continuar a ser possível associá-los à dimensão tempo, nem que seja com um valor pré-definido. Por isso, utiliza-se normalmente a criação de uma chave específica para a dimensão tempo, independente da data a que se referem os dados e que é referenciada na tabela de factos para corporizar a associação à dimensão tempo.

As problemáticas associadas à modelação da dimensão tempo variam em nível de dificuldade e complexidade na sua resolução, mas podem ser identificados alguns tipos de situações padrão, conforme as questões que se pretendem ver respondidas através da utilização de dados orientados temporalmente.

As questões mais simples, relacionadas com a selecção de conjuntos de registos da tabela de factos abrangidos por um período de tempo (por exemplo, numa determinada semana, mês ou conjunto de dias), podem ser facilmente respondidas através da associação da dimensão

tempo à tabela de factos, conforme já descrito.

No caso de se pretender obter a resposta a questões relacionadas com períodos temporais, mas que estão pré-definidos (por exemplo, meses, trimestres, anos) e onde não existe sobreposição de períodos, poderá ser utilizada uma dimensão tempo na qual seja indicado qual o período e o seu tipo(mês, trimestre) a que os dados dizem respeito.

Existem no entanto situações em que podem existir várias datas ou estampilhas temporais, associadas a um único registo da tabela de factos (por exemplo, a data de produção, data de venda, data de cobrança), mas que estão de alguma forma relacionadas. Neste caso, é necessário identificar univocamente o significado de cada uma, de forma a que o utilizador consiga seleccionar o contexto que lhe é mais interessante. Este será um caso em que poderão haver várias associações da tabela de factos à dimensão tempo, sempre que seja interessante a utilização de hierarquias associadas às várias datas disponíveis.

Uma situação diferente será, por exemplo, a obtenção de resposta a questões relacionadas com períodos temporais, arbitrários, que por vezes até se podem sobrepor, e que podem abranger apenas segundos, ou até meses. Uma forma de simplificar este tipo de pesquisas é associar a cada registo da tabela de factos duas estampilhas temporais, para assinalar quando a medida da tabela de factos passou a ter aquele valor, e outra para assinalar a data em que o valor foi novamente alterado. Assim, a pesquisa de um valor dentro de um determinado período de tempo resume-se a uma comparação de datas de início e fim.

Esta abordagem tem no entanto a desvantagem de que para cada registo da tabela de factos existem dois momentos de inserção: um quando o registo é criado, onde é preenchida a data de início, e outro para preencher a data de fim do período abrangido pelo registo.

Outra situação ainda mais complexa é a questão da marcação de tempo ao segundo. As questões que se põem aqui são, de uma forma geral, semelhantes às já descritas, mas neste caso temos os limites do período de tempo definidos ao segundo. Neste cenário, pode-se colocar as mesmas duas estampilhas temporais na tabela de factos, mas tem de se desistir da associação à dimensão tempo completa, pois normalmente não é praticável criar uma única dimensão com todos os minutos e segundos ao longo de um período de tempo. Pode-se, no entanto, criar chaves adicionais na tabela de factos, para que se consiga ligar uma dimensão tempo ao dia, por exemplo, para permitir a navegação dos dados, ficando o detalhe ao segundo nas estampilhas temporais.

2.3.2 Dimensões lentamente alteráveis

Tipicamente os valores das dimensões são conhecidos à partida, podendo ser acrescentados novos registos a cada dimensão, mas que não interferem nos valores numéricos já existentes (os novos registos da dimensão apenas serão utilizados em novos dados).

Existem no entanto algumas dimensões cujos valores se alteram ao longo do tempo, sendo que pode ser necessário a cada momento obter o(s) valor(es) da dimensão que estava(m) activo(s) no período de tempo para o qual se pretende efectuar a análise de dados. Como exemplos destas dimensões referem-se o "Produto" (alteração da categoria a que o produto pertence, renomeação de produtos, etc.) e "Cliente" (alteração da morada, estado civil, etc.).

São consideradas 3 formas principais de se lidar com as dimensões lentamente alteráveis, ou Slowly Changing Dimensions (SCD) [Kim96b]:

1. sobreposição do valor da dimensão - esta opção não mantém histórico, deve ser utilizada quando o anterior valor não é interessante para a realização de análises, como seja, por exemplo, para correcção de erros ou actualização de atributos descritivos (alteração de e-mail, nome da empresa, etc.);

2. criação de novo registo na dimensão - este é o tipo mais comum de solução para as SCD. Neste caso, se cada registo for identificado por um código específico, a chave do DW poderá por exemplo prever mais alguns dígitos, para permitir as novas versões do mesmo registo, e evitando assim as estampilhas temporais. Esta opção tem a vantagem de permitir a partição automática dos dados, mas é necessário ter em conta que haverá necessariamente um crescimento do número de registos da tabela de dimensão, que pode ou não ser relevante;

3. adicionar à dimensão um campo que indica o valor corrente - nos casos em que o valor antigo é legítimo, e é necessário mantê-lo e utilizá-lo, pode ser adoptada esta solução. Consiste em criar um campo onde é indicado o valor corrente, passando outro campo a indicar o valor antigo. Apenas as alterações mais recentes são guardadas, mas permite a visualização dos dados por cada uma das alternativas. Pode ser alargada para as 2 alterações mais recentes, mas nesse caso poderá ser considerada a adopção da segunda forma de se lidar com as SCD em vez desta.

   Esta última opção tem algumas desvantagens: é necessário um atributo extra por cada alteração que se pretenda guardar; caso se pretenda guardar a data de alteração, é também necessário contabilizar mais este atributo; e só permite guardar um número limitado de alterações. O código necessário para extrair a informação pode ser complexo, quando se quer o valor num período específico do tempo, ou vários atributos em simultâneo, com valores velhos e novos.

O primeiro passo para gerir as SCD é determinar qual o tipo de alterações que se vai acomodar, e a partir daí determinar a implementação pretendida.

Para as opções 1 e 3 não é necessária uma modificação substantiva no modelo de dados. Mas, no caso da opção 3, é necessário prever a inclusão dos novos campos para guardar as

alterações.

Quando se implementam dimensões agregadas, por exemplo para melhorar algumas pesquisas, é necessário ter em conta que as alterações de tipo 1 podem alterar a história e portanto têm influência ao nível de histórico das tabelas agregadas.

A implementação da opção 2 é mais complexa. Para esta opção de implementação torna-se particularmente importante (e é normalmente aconselhado) a utilização de chaves artificiais no DW, que não tenham qualquer significado a nível de negócio [Kim98]. Desta forma é sempre possível criar novos registos sem se comprometer a identificação da sua inter-relação (mantém-se o identificador interno do registo, apenas muda a chave).

Como forma alternativa de implementar esta opção podem ser utilizadas chaves compostas (acrescenta-se à chave existente mais 2 ou 3 dígitos para a versão), mas tem como desvantagem aumentar o tamanho da chave que depois será usada nas tabelas de factos.

No caso de se tratar de modelos normalizados, que normalmente não têm normalmente uma dimensão tempo explícita, a existência de um atributo que indique qual o valor corrente facilita a realização das pesquisas.

A implementação desta opção tem, portanto, custos de complexidade associados à alteração de chaves primárias. É necessário ter em conta que num sistema normalizado, ou dimensional mas com estrutura *snowflake*, a alteração da chave primária de uma tabela de dimensão pode implicar a actualização das chaves estrangeiras dos registos das tabelas relacionadas (por exemplo, quando se altera a chave primária num nível superior de uma hierarquia dimensional).

Assim, para se diminuir a complexidade, podem ser utilizadas soluções híbridas, por exemplo, do tipo 1 e 2, de forma a diminuir o número de campos abrangidos pela opção de implementação do tipo 2.

Todas estas abordagens permitem que as dimensões se mantenham em tabelas desnormalizadas, facilitando a sua utilização e navegação por todos os atributos da dimensão nas ferramentas de pesquisa. Além disso, torna transparente a existência de várias versões de um só registo, pelo menos até que seja feita uma selecção por um dos atributos alterados.

No entanto, pode acontecer que um atributo de uma destas dimensões se comece a alterar mais frequente, como por exemplo mensalmente. Nesse caso, deixamos de estar perante uma dimensão lentamente alterável [Kim99]. Uma das abordagens para solucionar este problema será a separação dos atributos problemáticos numa ou mais dimensões, criando mini-dimensões, e deixando os atributos lentamente alteráveis na dimensão maior, mas essa problemática não será detalhada neste trabalho.

2.4 Processos de Extracção, Transformação e Carregamento (ETL)

As organizações estão conscientes de que têm dados valiosos distribuídos ao longo das suas redes de informação que precisam de ser movimentados de um lado para o outro, como por exemplo entre diferentes aplicações de negócio, ou para um DW para efeitos de análise [Son04].

O único problema é que os dados estão em vários sistemas heterogéneos e portanto em diversos formatos. Para resolver este problema, as organizações utilizam *software* de ETL, que inclui a leitura dos dados na fonte, a sua limpeza e uniformização e a sua inserção no repositório final.

Os dados utilizados nos processos ETL podem ser provenientes de qualquer fonte, como por exemplo aplicações ERP [Wika], ficheiros de dados, folhas de cálculo. Após a extracção, os dados são transformados, ou modificados, dependendo da lógica de negócio envolvida, de forma a que possam ser enviados para o repositório de destino.

Existem diversas transformações que podem ser necessárias. A maioria das operações de transformação envolvem a limpeza dos dados para remover duplicados e forçar a consistência. Podem também envolver a normalização de nomes, endereços ou a expansão de registos com campos adicionais para armazenar informação demográfica ou dados provenientes de outros sistemas.

Existem duas abordagens para a implementação do ETL: utilizando ferramentas já existentes, ou codificando manualmente todo o processo ETL.

A primeira abordagem tem como vantagens, entre outras, o facto de ser mais simples, fácil e barata, se forem projectos complexos e com dimensão suficiente que justifiquem os custos de aquisição da ferramenta. Além disso, estas ferramentas de ETL incluem já os repositórios de metadados, sendo capazes de gerar automaticamente a informação para os preencher e por vezes podem até ser utilizadas por pessoas sem competências de programação.

Por outro lado, a abordagem da codificação manual permite flexibilidade ilimitada, pois pode ser desenhado tudo aquilo que se desejar. Não se fica limitado a um fornecedor nem às capacidades de uma única ferramenta (como por exemplo, a sua linguagem específica), e pode-se gerir directamente os metadados nos sistemas desenvolvidos à medida, se bem que é necessário nestes casos criar todos os interfaces de metadados.

Os processos ETL precisam de cumprir um certo conjunto de requisitos para que sejam robustos, resistentes a falhas, verdadeiramente úteis e completos. Apresenta-se de forma resumida alguns desses requisitos, seleccionados devido à sua relação directa com as problemáticas abordadas no âmbito deste trabalho [KC04].

- Devem garantir a qualidade, mantendo um registo completo de auditoria associado ao facto ou dimensão final. Os dados que se apresentem corrompidos ou suspeitos precisam

de ser tratados através de um número definido e limitado de respostas, como por exemplo o preenchimento de texto em falta através de um símbolo específico, ou obter estimativas de valores numéricos, que podem ainda existir, mas foram corrompidos antes de serem transferidos para o DW.

- Após ocorrer uma interrupção anormal os processos ETL devem conseguir efectuar a recuperação (*recovery* no original inglês), e recomeçarem a partir do ponto exacto onde tinham ficado, para que não hajam actualizações repetidas ou em falta.
- Os metadados provenientes dos Sistemas de Gestão de Bases de Dados (SGBD) e das ferramentas de desenho são fáceis de capturar, mas geralmente constituem apenas uma pequena parte dos metadados necessários para perceber e controlar completamente o sistema. Outra parte dos metadados é gerada pelos processos de limpeza, mas o maior desafio que se põe à equipa de ETL é como e onde se guarda a informação relativa aos fluxos dos processos. Tipicamente as soluções de ETL mantêm esta informação automaticamente, conforme já referido; se os processos de ETL tiverem sido desenvolvidos à medida, é necessário implementar um repositório central desta informação.
- Uma abordagem sistemática à segurança exige que sejam adoptadas medidas físicas e administrativas que cubram cada tabela e cada cópia do ambiente ETL. Os dados devem ser arquivados com formas de verificação (por exemplo, *checksum*) que demonstrem que não foram alterados. Os dados mais sensíveis precisam de ter associados instrumentos que registem cada acesso ou comando efectuado por cada administrador, e que possam ser consultados.

Ao longo desta secção serão abordadas com maior detalhe a problemática da qualidade dos dados e, directamente relacionada com esta, a rastreabilidade dos dados.

Dada a sua relevância, complexidade e quantidade de informação disponível, os metadados são apresentados numa secção própria (2.5).

A problemática da segurança não é estudada especificamente no âmbito deste trabalho.

2.4.1 Qualidade dos Dados

Os três maiores problemas relacionados com a qualidade dos dados podem ser descritos como [Kim96a]:

1. acesso aos dados - existe uma grande quantidade de informação, mas que não está directamente acessível;
2. pesquisas - o sistema deve mostrar o que é importante, mas o utilizador tem de conseguir identificar o que está a ver, para poder utilizar esses resultados;
3. integridade dos dados - alguns dos dados podem não ter muita qualidade. Por exemplo,

se não existirem listas centralizadas de clientes, será de esperar situações de redundância e mesmo erros na informação sobre clientes.

Os primeiros dois problemas já foram abordados, tanto ao nível empresarial como técnico. Para a resolução do primeiro problema contribuiu o desenvolvimento do *hardware* e *software* de DW, para permitir a integração da informação disponível e rápido acesso à mesma. Para resolver o segundo problema é possível utilizar esquemas de interligação dos utilizadores aos dados, que recorrem ao auxílio de metadados para documentação da informação. No entanto, para resolver o terceiro problema, a integridade dos dados, não foi ainda possível definir uma abordagem única que garanta a sua resolução, pois a falta de qualidade dos dados pode ser imputada a vários factores, como sistemas de recolha de informação implementados de forma deficiente, ou más práticas na captura dos dados.

O processo de limpeza dos dados, que contribui para garantir a sua integridade, não consiste apenas na actualização de um registo com dados correctos. Pode envolver a decomposição e novamente junção dos dados. Este processo pode ser dividido em 5 passos:

- converter em elementos - através de um interpretador de informação, subdividir os dados nos seus elementos mais simples;
- normalizar - colocar todos os elementos comuns representados da mesma forma (por exemplo, nas moradas, colocar todos os "R.´´ como "Rua´´);
- verificar - a normalização efectuada falhou nalgum caso? Nem todos os elementos que se julgou terem o mesmo significado o tinham na realidade?
- encontrar correspondências - verificar se noutro conjunto de dados existem alguns elementos comuns que por si só ajudem no processo de verificação, ou se os dados são realmente iguais (assinalando as alterações efectuadas);
- documentação - é necessário documentar os resultados das operações anteriores sob a forma de metadados. Isto ajudará a assegurar que as futuras limpezas de dados reconhecerão melhor os dados correctos, e que as aplicações conseguirão realizar melhor as operações de *slice and dice*[1], e compreender a base de dados.

Assim, a implementação de processos de limpeza de dados não é simples e torna-se muitas vezes difícil de praticar quando estamos perante grandes volumes de dados. A falta de qualidade dos dados pode, no entanto, ser evitada. A principal resposta para evitar o problema dos dados com falta de qualidade consiste na reengenharia dos processos de negócio, nas suas várias vertentes, tais como:

- providenciar sistemas de entrada de dados bem arquitectados, com o maior número de regras de validação possível;

[1] Operações que permitem aceder a um DW através de qualquer uma das suas dimensões de igual forma, ou seja, permitem navegar através dos dados do cubo de decisão ao longo de qualquer dimensão. [Mor]

- criação de conjuntos únicos de regras de negócio para validação de dados para cada um dos tipo de dados;
- ao nível da gestão executiva, assegurar que a entrada de dados com qualidade tem prioridade máxima;
- fomentar uma cultura na empresa que valorize o detalhe e a qualidade dos dados.

Adicionalmente, se para cada dado for possível identificar univocamente a sua origem, será mais fácil corrigir eventuais deficiências que sejam detectadas. E, será mais fácil para os utilizadores compreenderem os dados que serviram de base, evitando dúvidas e desconfiança relativamente aos resultados. Esse é o principal papel do rastreio dos dados (ver secção 2.4.2).

No entanto, para além da identificação da origem dos dados, podem ser introduzidos descritores de controlo de qualidade associados aos dados, para auxiliar no processo de avaliação da sua qualidade. São exemplos destes descritores o dia/hora em que a extracção de dados ficou completa, a data de fim em que as correspondências foram encontradas, a data da geração da chave do DW, a data da criação ou actualização das agregações, a data da última correcção efectuada, a identificação unívoca dos processos ETL utilizados no carregamento de dados (se os metadados ETL forem correctamente administrados, será possível reconstituir pormenorizadamente os processos utilizados), etc.

Podem também ser criadas métricas quantitativas para avaliação da qualidade dos dados [Kim]:

- completude do conjunto de dados relativamente ao máximo possível;
- número de elementos de dados que fundamentam o resultado observado;
- número de elementos de dados não aplicáveis encontrados no conjunto de entrada;
- número de elementos de dados corrompidos encontrados no conjunto de entrada;
- número de elementos de dados fora dos limites definidos encontrados no conjunto de entrada;
- número de elementos desconhecidos tratados como zero;
- número de elementos de dados alterados manualmente no processo de carregamento;
- número de elementos de dados não classificados nos agregados regulares;
- número de correcções efectuadas desde o carregamento original.

Uma opção para implementar estes indicadores de qualidade é associá-los directamente aos dados, criando uma dimensão ”Auditoria´´ ligada ao registo do dado no seu maior nível de detalhe. Naturalmente, para que uma dimensão deste tipo seja útil, é necessário que os processos ETL tenham rotinas que façam o correcto seguimento de cada dado que for carregado. E

também é necessário que todas as descrições e datas relevantes sejam correctamente registadas, porventura tornando mais complexos os próprios processos ETL.

2.4.2 Rastreabilidade dos dados

O rastreio da informação, é um conceito associado à capacidade de identificar a história, aplicação e localização de um item ou actividade através dos registos de informação.

A gestão do rastreio de dados implica a recolha, armazenamento, processamento e disponibilização de grandes quantidades de informação ao longo do ciclo de vida dos itens ou actividades, que devem estar permanentemente disponíveis para todos os agentes envolvidos no processo, desde o início do processo.

As características básicas do rastreio dos dados são: identificação das origens de todos os componentes envolvidos, informação sobre quando e para onde foram movidos ou transformados (rastreio dos produtos), e um sistema que interligue toda a informação (rastreio dos dados).

A possibilidade de se efectuar o rastreio dos dados relativos aos produtos e actividades tem tido especial desenvolvimento e impacto em áreas críticas, onde a necessidade de identificar rigorosamente a origem e as transformações sofridas pelos produtos são essenciais, como por exemplo as indústrias alimentares (que envolvem um grande risco de negócio, onde um único ingrediente, ou um só pacote contaminado, é um risco real para a saúde pública) [dCNRPB03], e laboratórios de análises (onde é preciso identificar univocamente a origem dos resultados obtidos).

Os requisitos de rastreabilidade são essenciais para se perceber os efeitos em cascata das alterações propostas aos produtos resultantes das tarefas de desenvolvimento, evitando assim que os produtos finais fiquem inconsistentes quando se fazem alterações. Este aspecto tem particular relevância quando numa abordagem incremental se pretende uma definição evolutiva dos requisitos de sistema. Sempre que novos requisitos de sistema são definidos, eles devem ser rastreados nas futuras versões definidas pelo plano de desenvolvimento.

Os requisitos de rastreabilidade devem ser implementados de forma a que, sempre que qualquer requisito seja alterado, todos os requisitos relacionados, os componentes implicados e os casos de teste afectados pela alteração sejam identificados.

Tipicamente os DW têm como objectivo receber e agregar dados provenientes de várias fontes, para possibilitar a realização de análises para suporte à decisão. Para realizar a migração dos dados dos sistemas fonte para o DW existe um conjunto de processos ETL, através do qual os dados sofrem derivações e transformações para que possam ser mais eficientemente utilizados para a realização de análises de suporte à decisão. Assim, os dados apresentados no DW na maioria das vezes estão num formato completamente diferente do observado no

Sistema Operacional (SO), encontrando-se muitas vezes com níveis de agregação superiores ou sob formas pré-calculadas.

No entanto, torna-se muitas vezes necessário ao utilizador que realiza as análises no DW, a partir dos dados agregados (menor detalhe), descer ao nível de maior detalhe, para verificar a informação que lhe está a ser apresentada.

Pode ainda ter necessidade de associar o conjunto de dados que estão a ser consultados no DW à sua origem no SO fonte, para melhor analisar o seu significado. O próprio pessoal de Tecnologia de Informação (TI) responsável pela manutenção do DW tem de ter esta possibilidade, já que uma utilização muito comum do mapeamento entre os dados fonte e destino é a verificação da origem de dados suspeitos ou anómalos.

Esta necessidade é descrita como o problema da linhagem dos dados.

Os analistas podem utilizar a linhagem dos dados para autenticar os dados fonte, bem como as regras de transformação utilizadas no Data Warehouse. É também uma maneira de promover a confiança dos utilizadores nos dados apresentados, já que eles terão a qualquer momento a possibilidade de confrontar os dados dos SO com os do DW. Inclusive, se a linhagem dos dados existir e for completa, facilita as análises de impacto no DW de eventuais transformações ou actualizações efectuadas no SO.

Em alguns sistemas, como por exemplo o WHIPS [CW00], esta questão foi resolvida através do desenvolvimento de instrumentos de registo do conjunto de derivações associadas desde a fonte (SO) até ao destino (DW). Quando se pretende encontrar os dados de origem, inverte-se o processo das derivações, conseguindo-se assim obter o caminho percorrido pelos valores.

A solução para o problema da linhagem dos dados é constituída por 2 passos [Var02]:

1. Desenhar e construir uma boa estratégia para registar a linhagem dos dados.
2. Conseguir utilizar efectivamente a informação registada para determinar a linhagem dos dados.

Relativamente ao primeiro ponto, foram já propostos vários métodos para capturar informação de linhagem dos dados. A maioria dos modelos comerciais utiliza os metadados resultantes do processo de ETL para rastrear os dados. A linhagem dos dados pode ser mantida ao nível da tabela, da coluna ou mesmo do registo de dados. O analista limita-se a utilizar os metadados ETL para determinar o caminho percorrido pelos dados enquanto foram transferidos dos sistemas fonte para o DW.

Um dos problemas desta abordagem é que devido à natureza dinâmica dos DW, os processos ETL vão sendo alterados ao longo do tempo, pelo que tem de se manter o controlo das versões anteriores, que reflectiam as transformações dos dados na data em que foram introduzidos no DW.

Até há pouco tempo, o CWM (Common Warehouse Metamodel) da OMG (Object Manage-

ment Group) e o OIM (Open Information Model) da MDC (MetaData Coalition) eram os dois *standards* mais influentes para a definição de metadados de DW. Em Setembro de 2001 o MDC terminou as operações e fundiu-se com a OMG para trabalhar num conjunto único de especificações de metadados (ver capítulo 2.5). A linhagem dos dados continuará a ser um elemento importante do standard integrado.

Uma outra abordagem, particularmente útil quando a ferramenta ETL utilizada não suporta directamente o registo da linhagem dos dados, e que pode ser utilizada, é bastante simples.

Primeiro adiciona-se uma coluna a cada uma das tabelas do DW que registará a data e hora do carregamento do registo para o DW. A seguir, adiciona-se a cada uma das tabelas do repositório onde são registados os metadados provenientes do processo ETL uma data de início e de fim. De cada vez que os processos ETL forem alterados, as datas de fim destes registos terão de ser actualizadas e inseridos novos registos no repositório, reflectindo todas as alterações efectuadas aos processos. Assim, as datas de início e fim indicarão os processos ETL utilizados para carregar os registos para o DW, através da verificação da data de carregamento.

Mas, quando se chega ao ponto de utilização da informação, é necessário ter em conta que estas técnicas só permitem identificar os campos de origem dos dados no SO e quais as transformações que sofreram, não dizem por extenso o conjunto de dados que foi utilizado para o cálculo dos valores que estão a ser observados.

O utilizador pode sempre tentar efectuar uma inversão das transformações sofridas pelos dados para chegar à fonte, mas dada a complexidade da tecnologia actual, é muito provável que para efectuar esta operação seja necessário o auxílio de um analista programador.

Se for possível navegar desde o DW até aos dados fonte, mesmo quando existam várias transformações de diferentes complexidades no processo ETL, pode-se dizer que se completou a rastreabilidade dos dados.

2.5 Metadados

Os metadados são descritos como informação estruturada que descreve, explica, localiza ou torna mais fácil obter, usar ou gerir um recurso de informação. São muitas vezes chamados ”Dados sobre dados´´, ou ”Informação sobre informação´´ [Org04].

Existem 3 tipos principais de metadados:

- Metadados descritivos - descrevem um recurso com o objectivo da descoberta e identificação, e podem incluir elementos como o título, resumo, autor e palavras-chave.
- Metadados estruturais - indicam como os objectos compostos são unidos. Por exemplo, como as páginas são ordenadas para formar capítulos.

- Metadados administrativos - fornecem informação para ajudar a gerir um recurso. Por exemplo, como e quando foi criado, tipo de ficheiro, controlo de acessos, e outra informação técnica.

Por vezes são distinguidos ainda outros 2 tipos de metadados:

- Metadados de gestão de direitos - relativos aos direitos de propriedade intelectual.
- Metadados de preservação - contêm informação necessária para arquivar e preservar um recurso.

Os metadados podem ser armazenados juntamente com o objecto que pretendem descrever, o que assegura que não serão perdidos e que serão actualizados simultaneamente com o objecto. Ou separadamente, o que pode simplificar a gestão dos metadados, e facilitar a pesquisa e obtenção de informação.

Muitas vezes os metadados são armazenados em documentos HTML [Conb] ou XML [Cona], mas é bastante mais eficiente e cada vez mais comum que as páginas sejam construídas dinamicamente a partir dos metadados armazenados, em bases de dados, ou qualquer outro formato *standard*.

A descrição de um recurso com metadados permite que ele seja totalmente compreendido, quer por humanos quer por máquinas, promovendo a interoperabilidade. Esta é por definição a capacidade de múltiplos sistemas, com diferentes plataformas *hardware* e *software*, estruturas de dados e interfaces, trocarem dados entre si com uma perda mínima do conteúdo e funcionalidade.

Os metadados são também a chave para assegurar que os recursos continuarão a ser acessíveis no futuro. O arquivo e preservação requerem elementos especiais para rastrear os dados (ver capítulo2.4.2) de um determinado objecto, de forma a detalhar as suas características físicas e a documentar o seu comportamento, para o poder simular com novas tecnologias.

Os metadados podem ser estruturados sob a forma de esquemas, compostos por conjuntos de elementos desenhados para um objectivo especifico (por exemplo, descrever um tipo de recurso de informação), e que especificam os nomes e a semântica dos seus elementos. A semântica de um esquema de metadados é construída a partir da definição ou significado dos seus elementos. Os valores dados aos elementos de metadados são os conteúdos. Os esquemas de metadados podem também especificar regras sobre como descrever o conteúdo, regras de representação do conteúdo e valores permitidos para os conteúdos. Também podem exprimir regras sintácticas, que descrevem como os elementos e o seu conteúdo devem ser codificados (se um esquema não tiver regras sintácticas definidas, diz-se que é independente da sintaxe).

Muitos esquemas de metadados actuais utilizam SGML (Standard Generalized Markup Language) [Conc] ou XML (eXtensible Markup Language) [Cona].

No âmbito deste trabalho foram estudados vários esquemas de metadados, como a Re-

source Description Framework (RDF) do W3C [Sni01], uma arquitectura desenhada para permitir a partilha de metadados e o Common Warehouse Metamodel (CWM), do Object Management Group (OMG). Dado que o CWM pretende ser um *standard* para descrição, tradução e troca de metadados, optou-se por lhe dar maior relevância no âmbito deste trabalho, sendo apresentada na próxima secção uma descrição resumida deste esquema.

2.5.1 Common Warehouse Metamodel (CWM)

No final de Agosto de 2000, os dois standards de metadados CWM (Common Warehouse Metamodel) da OMG (inicialmente proposto pela IBM, Oracle e Unisys) e o OIM (Meta Data Coalition's Object Information Model) da Microsoft iniciaram a sua união, com o OIM a ser incorporado no CWM. Por esse motivo, o OIM não será apresentado nesta dissertação, mas apenas será apresentado um resumo do CWM.

O CWM [PCTM03] utiliza UML, XML e XML Metadata Interchange (XMI) [Wikd], definindo um metamodelo comum para a armazenagem de dados e normaliza a sintaxe e semântica necessárias para a importação, exportação e outras operações dinâmicas do DW.

A sua especificação inclui Application Programming Interfaces (API), formatos de troca, e serviços que suportam todo o ciclo de vida da gestão de metadados incluindo a extracção, tranformação, transporte, carregamento, integração e análise.

O metamodelo do CWM consiste num conjunto de sub-metamodelos, ou pacotes, que representam os metadados comuns das principais áreas de interesse do *data warehouse* e *business intelligence*, divididos por vários níveis, conforme se apresenta na figura 2.4.

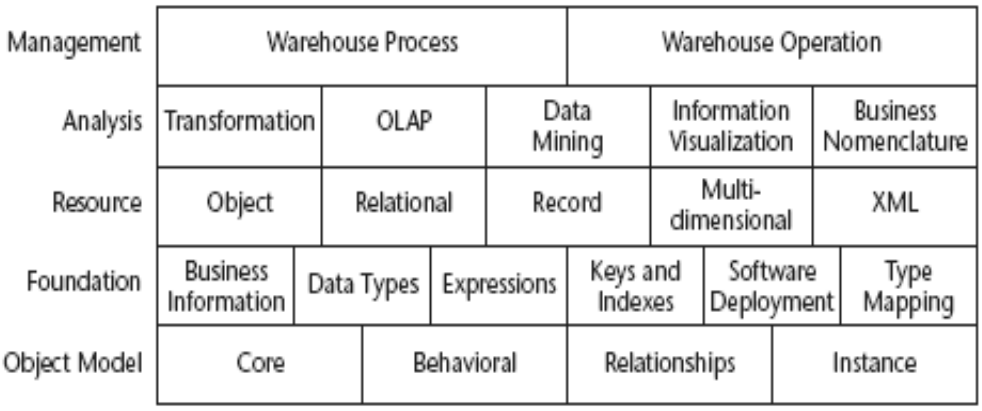

Figura 2.4: Metamodelo CWM: estrutura de pacotes

- Modelo de objectos - os pacotes deste nível permitem definir os conceitos fundamentais dos metamodelos, as relações e as restrições necessárias para os restantes pacotes CWM.
- Fundações - providencia serviços específicos de CWM para pacotes dos níveis superiores.

- Recursos - descreve as estruturas dos recursos de dados que vão agir como fontes ou destinos da comunicação através do CWM. Inclui metamodelos que representam recursos de dados relacionais, orientados por objectos, multidimensionais e XML.
- Análise - descreve serviços que operam nos recursos de dados de origem e destino definidos no nível "Recursos´´. Permite representar transformações de dados, OLAP, *data mining*, visualização da informação e nomenclatura de negócio.
- Gestão de armazém - providencia funções e serviços de suporte da operação diária e gestão de DW. Permite representar processos de armazém e resultados de operações.

Nesta dissertação aprofundaremos o estudo do CWM apenas relativamente a dois sub-metamodelos - o multidimensional e o OLAP, dado que são os mais relevantes em termos da compreensão conceptual do modelo apresentado na presente dissertação.

2.5.1.1 Metamodelo multidimensional

O metamodelo multidimensional [CCC+01], apresentado na figura 2.5, é uma representação genérica de uma base de dados multidimensional, não tentando fornecer uma representação completa de todos os aspectos das bases de dados multidimensionais disponíveis comercialmente, até porque estas tendem a ter estruturas proprietárias e não existem representações padrão amplamente aceites dos seus esquemas lógicos.

O *Schema* inclui todos os elementos do modelo multidimensional e representa a instância da base de dados multidimensional considerada. A *Dimension* representa uma dimensão física na base de dados multidimensional.

Enquanto o metamodelo OLAP (apresentado a seguir neste capítulo) define uma dimensão como uma entidade puramente conceptual, no metamodelo multidimensional a *Dimension* representa o objecto dimensão, criado pelo modelo de programação da base de dados multidimensional. Uma *Dimension* pode ainda incluir outras instâncias de *Dimension*, para formar estruturas dimensionais complexas, como por exemplo, hierarquias com diferentes níveis de detalhe (por exemplo, como foi apresentado no esquema *snowflake*).

Os *DimensionedObjects* representam os atributos da *Dimension* e encontram-se incluídos no *Schema*, sendo referenciados pelas *Dimension* que os utilizam. Estes objectos incluem medidas (também chamadas variáveis ou métricas de cálculo), fórmulas, funções de consolidação, sinónimos dos nomes dos membros, etc.

O *MemberSet* representa a colecção de membros associados a uma instância de *Dimension*, o *MemberValue* representa uma instância do valor de um *Member*.

O *MemberSet*, *Member* e *MemberValue* possibilitam a interligação entre os níveis M1 (nível correspondente ao modelo e seus metadados) e os respectivos valores do nível M0 (nível dos

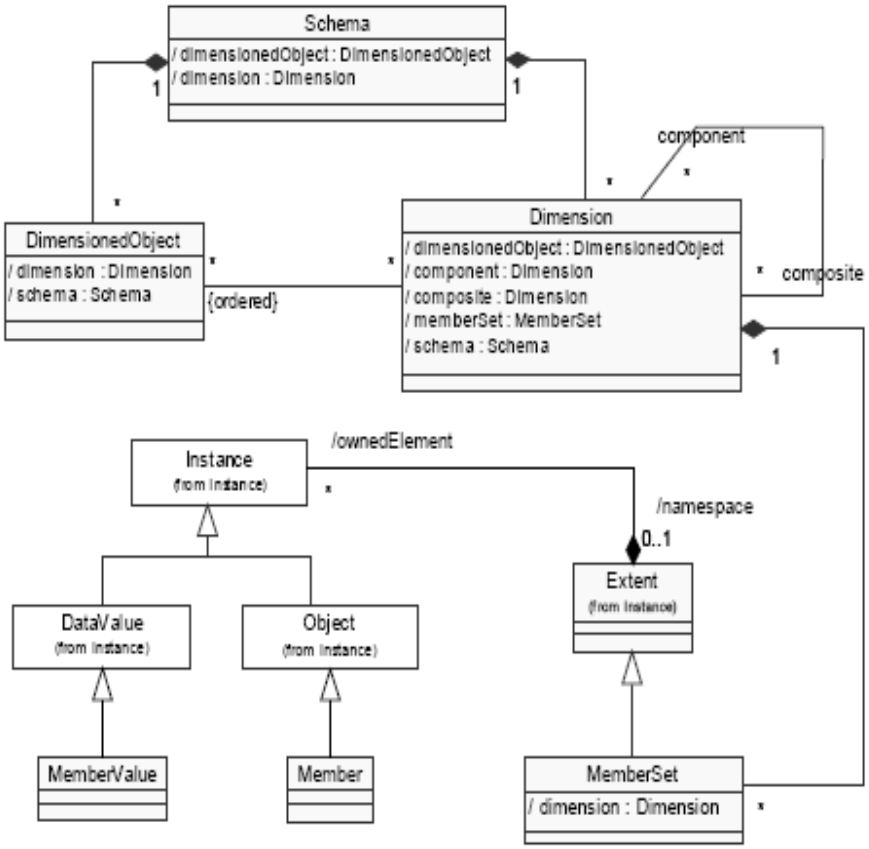

Figura 2.5: Metamodelo multidimensional CWM: classes e associações

objectos e dados).

2.5.1.2 Metamodelo OLAP

Os principais objectivos do *package* CWM *OLAP* (apresentado na figura 2.6) são:

- Definir um metamodelos de conceitos essenciais de OLAP, comuns à maioria dos sistemas OLAP.
- Providenciar uma estrutura onde as instâncias do metamodelo OLAP são mapeadas para estruturas capazes de suportar a entrada em exploração (modelos de recursos físicos de dados).
- Assegurar que a navegação através da hierarquia do modelo OLAP e os seus vários modelos de recursos é sempre efectuada de uma forma uniforme.
- Nivelar os serviços providenciados por outros pacotes CWM.

No metamodelo OLAP, o *Schema* contém *Dimensions* e *Cubes* e é o contentor lógico de todos os elementos que o constituem. É o elemento raiz da hierarquia do modelo, marcando o ponto de entrada para navegar o modelo OLAP.

Uma *Dimension* consiste numa lista de valores únicos (chamados membros) que partilham um significado comum dentro do domínio que está a ser modelado.

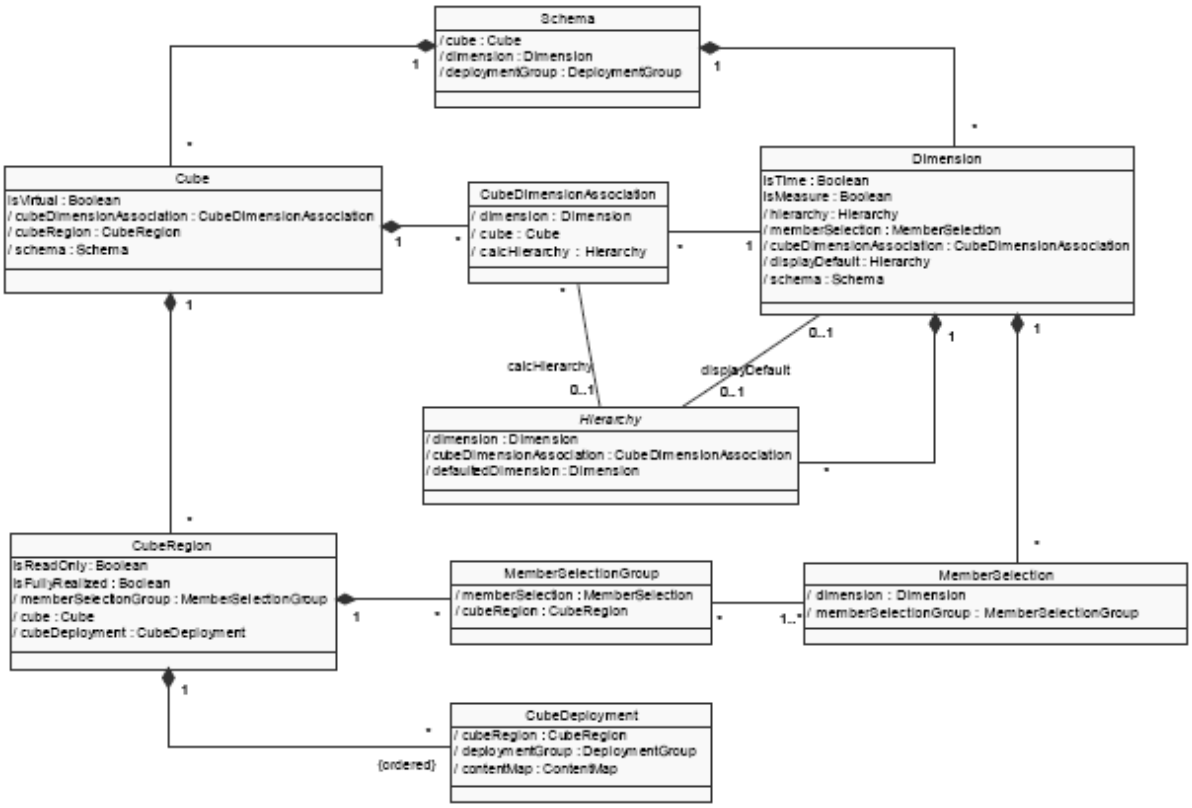

Figura 2.6: Metamodelo OLAP CWM: classes e associações

Um *Cube* é uma colecção de valores analíticos (medidas) que partilham a mesma dimensionalidade, especificada por um conjunto de dimensões únicas pertencentes ao *Schema*. Cada combinação única de membros no produto cartesiano das dimensões do *Cube* identifica univocamente uma célula de dados dentro da estrutura multidimensional.

A *CubeDimensionAssociation* relaciona um *Cube* com as dimensões que o constituem, e serve para expor características relevantes nas relações *Cube-Dimension*, como por exemplo as hierarquias de cálculo.

Uma hierarquia é uma estrutura organizacional que descreve um padrão transversal a uma dimensão, baseado em relações pai/filho entre os membros da dimensão. Assim, uma dimensão pode ter uma ou mais hierarquias, sendo estas usadas para definir os caminhos de navegação e de consolidação através da dimensão.

Pode acontecer que um determinado membro seja agregado por mais do que um 'pai'. Por exemplo, uma dimensão tempo que no seu nível base seja baseada em dias pode ter uma hierarquia que especifique a consolidação dos dias em semanas, e estas em anos, e outra que agregue os dias em meses, estes em trimestres e finalmente os trimestres em anos. Normalmente é sempre definida uma hierarquia por omissão para os efeitos de cálculos de agregação efectuados no *Cube*.

Os mecanismos de *MemberSelection* permitem particionar a colecção de membros da *Dimension*, ou seja, cada *MemberSelection* pode definir um subconjunto de membros.

As *CubeRegion* são utilizadas para implementar os *Cube*, ou seja, um *Cube* pode ser realizado a partir de um conjunto de *CubeRegion* que mapeiam partes do cubo para fontes de dados físicas. Os *MemberSelection* que definem as *CubeRegion* também podem ser agrupados através de *MemberSelectionGroup*, permitindo que as *CubeRegion* sejam definidas com uma semântica especifica.

Uma *CubeRegion* pode conter um número indeterminado de *CubeDeployment*, uma meta-classe que representa uma estratégia de implementação para uma estrutura multidimensional.

2.5.1.3 Utilização actual do CWM

Sendo o objectivo do CWM, representado na figura 2.7, bastante ambicioso, tentou-se neste trabalho de tese incorporar alguns exemplos da utilização efectiva do CWM como metamodelo para a troca e partilha de metadados, mas no entanto não foi possível encontrar publicações ou outros elementos sobre esta utilização.

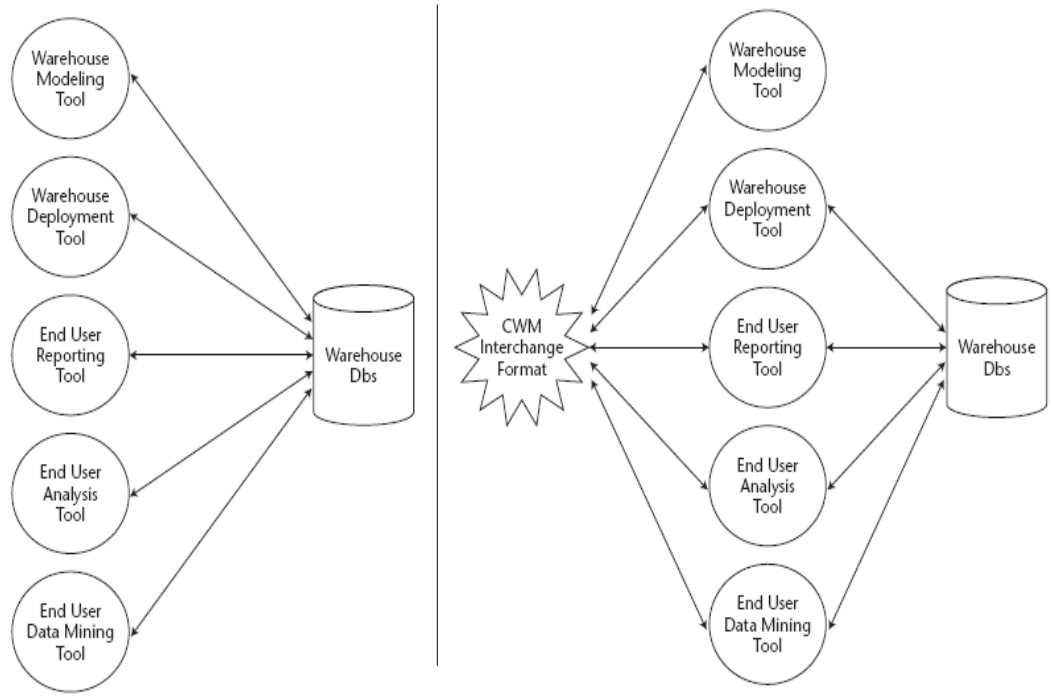

Figura 2.7: Apresentação de um grupo de ferramentas de Data Warehouse com um repositório de informação partilhado (à esquerda), e do objectivo pretendido com o CWM, possibilitando troca de informação entre elas (à direita)

O próprio OMG parece não estar muito activo no desenvolvimento do CWM, pois os últimos documentos encontrados na página do OMG que estão relacionados com o CWM têm data de 2001. A excepção é um documento de especificação lançado pelo OMG, o "CWM Metadata Interchange Patterns Specification", que data de Março de 2004, e onde participam

a Hyperion Solutions Corporation®, a Oracle Corporation®e a Unisys Corporation®(3 dos 8 participantes iniciais). Existem algumas implementações efectivas do CWM em ferramentas comerciais, como por exemplo a ferramenta Warehouse Builder™da Oracle®, que permite a descrição dos modelos multidimensionais construídos num formato XML baseado no *standard* CWM. Mas não é possível encontrar documentação que comprove a efectiva aceitação do CWM, enquanto norma para troca de metadados, no conjunto generalizado das ferramentas comerciais de DW existentes e a utilização deste metamodelo em sistemas efectivamente implementados e em produção.

2.6 Modelação Conceptual - YAM^2

Foram consideradas várias possibilidades para modelos conceptuais a serem utilizados no âmbito deste trabalho, como por exemplo a Modelação Dimensional, uma técnica bastante utilizada em DW e associada a Kimball e a modelação convencional E-R, associada a Inmon [Fir98].

Ambas as abordagens, estando interligadas com os modelos das estruturas de dados que suportam o DW, são relativamente limitadas em termos da expressividade dos modelos que permitem construir, sendo que expressividade de um modelo pode ser definida como o nível a que um modelo consegue representar, ou expressar, um conceito do mundo real. Ou seja, quanto mais expressivo for o modelo, melhor representará o mundo real e mais informação sobre os dados poderá ser retirada dele.

As aplicações OLAP requerem o máximo de expressividade possível, o que é uma necessidade crucial para os modelos conceptuais multidimensionais.

O YAM^2 [ASS02] surgiu em 2002, e é uma extensão ao UML [OMG] tal como o CWM, mas, embora o CWM permita a representação de conceitos multidimensionais, é muito geral e não foi concebido como um modelo conceptual, ao contrário do YAM^2.

Ao utilizar o UML como base para a definição das estruturas do YAM, este modelo fica assente numa base sólida e de uso generalizado, evitando assim que alguns conceitos básicos tenham de ser definidos. Por este motivo, o YAM^2 prevê a existência de vários tipos de nós e arcos nos grafos utilizados, que têm como objectivo aumentar a expressividade do modelo conceptual.

Um factor também importante para um modelo de dados é a capacidade que este tem de acomodar diferentes concepções, isto é, a sua "relatividade semântica".

A informação presente num DW deve ser apresentada aos utilizadores da forma como eles a compreendem melhor, independentemente de como foi originalmente concebida ou armazenada. Assim, o YAM^2 também providencia mecanismos - as relações de derivação - que

permitem modelar os mesmos dados sob diferentes pontos de vista.

Exactamente porque o YAM^2 é vocacionado para a modelação de modelos multidimensionais, este foi o método escolhido para a apresentação dos sub-modelos que corporizam o trabalho apresentado neste documento, embora tenham sido efectuadas algumas alterações à descrição original desta modelação conceptual e que serão apresentadas ao longo desta secção.

A base dos modelos multidimensionais é a dualidade factos-dimensões, em que os factos representam conjuntos de dados a serem analisados (representam "medições") e as dimensões mostram diferentes pontos de vista que podem ser utilizados nas análises. Sendo o YAM^2 uma extensão ao UML, utiliza o paradigam O-O ("Object Oriented"), pelo que se apresentam os tipos de nós que podem ser encontrados num sistema multidimensional orientado por objectos, e que são utilizados no YAM^2 para efectuar a representação conceptual dos modelos multidimensionais:

- Dimensão - grafo conexo dirigido e que representa um ponto de vista na análise de dados, ou um eixo de análise. Os nós do grafo representam níveis (uma dimensão pode ter um ou mais níveis), e os arcos representam as relações entre níveis, sendo que cada instância do nível de destino é composta por um conjunto de instâncias do nível de origem.
- Nível - representa um conjunto de instâncias com a mesma granularidade, numa dimensão de análise. Ou seja, todos os descritores presentes naquele nível reflectem o mesmo nível de detalhe dos dados.
- Descritor - atributo de um nível, usado para seleccionar algumas das suas instâncias. É uma especialização da metaclasse UML "Attribute´´.
- Facto - é um grafo conexo, dirigido, que representa um assunto de análise. Cada nó do grafo é uma célula. Cada ligação entre células reflecte que a instância da célula de destino (nível de maior agregação) é composta por um conjunto de instâncias da célula de origem (com menor agregação).
- Célula - é um conjunto de instâncias de um dado tipo de facto, medido à mesma granularidade para cada uma das suas dimensões de análise. É uma especialização da metaclasse UML "Class´´.
- Medida - é um atributo de uma célula representando os dados medidos e que vão ser analisados. Cada instância de uma célula contém um conjunto de medidas (que pode ser vazio). As várias medidas que se encontram em células diferentes correspondem ao mesmo conceito medido, mas com diferentes níveis de agregação, definidos pelas células. É uma instância da metaclasse UML "Atributo´´.

Estes 6 nós são agrupados aos pares, conforme o seu nível conceptual.

Note-se que a palavra "nível" assume neste contexto um duplo significado, enquanto con-

stituinte de uma dimensão e enquanto indicativo de nível do modelo conceptual - superior, intermédio ou inferior. Para diferenciar ambos os casos, o nível aplicado ao modelo conceptual será sempre designado por "nível conceptual".

O nível conceptual superior inclui os Factos e Dimensões, os constituintes de uma Estrela. Uma estrela é composta por um só facto e várias dimensões.

O nível conceptual intermédio inclui as Células e Níveis.

O nível conceptual inferior é o que representa a informação de maior detalhe e inclui as Medidas (que podem ser agrupadas por Tipo de Medidas, embora no âmbito deste trabalho esta possibilidade não seja utilizada) e os Descritores.

Apresenta-se na figura 2.8 a organização das cardinalidades entre as diferentes estruturas.

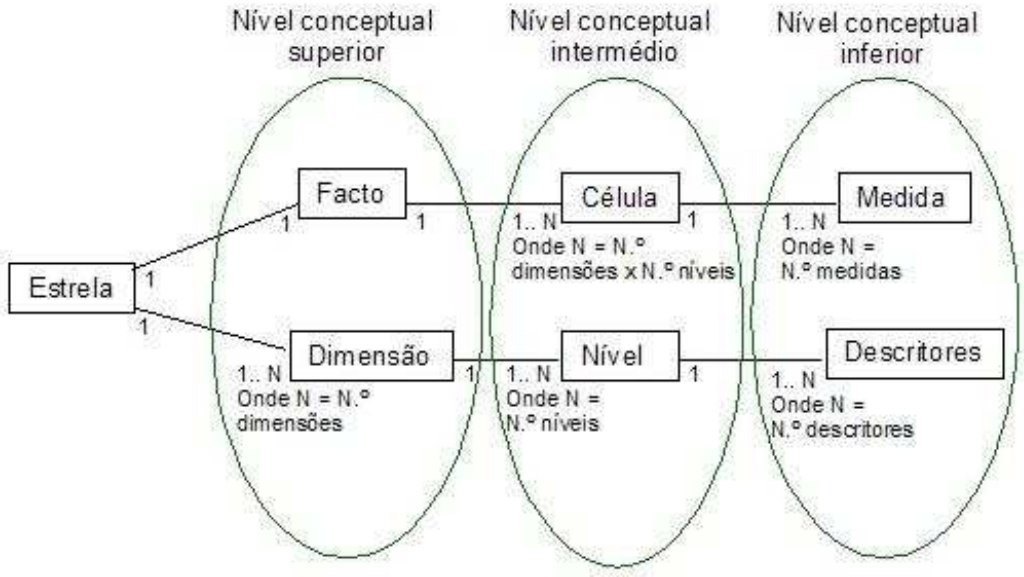

Figura 2.8: YAM: Estrutura dos nós YAM[2]

Uma Estrela contém apenas um único Facto e várias Dimensões. Cada Facto pode ser constituído por várias Células, cada Dimensão é constituída por vários Níveis. Cada Célula de um Facto pode ter várias Medidas e cada Nível de uma Dimensão pode ter vários Descritores.

Para melhor explicar o conceito de Dimensão, Nível e Descritor temos como exemplo a dimensão "Cliente", que inclui vários atributos descritivos de um cliente, como por exemplo o seu escalão etário. Uma abordagem de modelação desta dimensão poderá considerar 3 níveis: o nível de agregação máximo (todos os clientes), o nível intermédio, onde se encontra o escalão etário ou a classificação do cliente por idade e, o nível inferior ou de agregação mínimo, onde se instancia cada um dos clientes.

A representação YAM desta dimensão seria conforme se apresenta no lado direito da figura 2.9. No lado esquerdo, apresenta-se uma representação simples da estrutura da dimensão.

Os descritores do cliente poderiam ser, por exemplo o nome e morada no nível inferior e o seu escalão etário (por exemplo, com os valores 1 a 20 anos, 21 a 35 anos, 36 a 50, etc.) no nível

Figura 2.9: YAM^2: Exemplo de representação da dimensão "Cliente"

intermédio. Ou seja, os descritores permitem caracterizar quais os grupos de clientes que se desejam analisar, sendo que no limite se pretende considerar o próprio cliente (seleccionando por exemplo o nome).

No YAM^2 coloca-se uma letra (C) ou (D) para identificar nos diagramas as Células e as Dimensões, respectivamente. Neste trabalho não são utilizadas essas letras, sendo as Dimensões assinaladas com o sufixo"_Dim´´ e em vez de se indicarem "Células", é utilizada a expressão "Factos" (identificados pelo sufixo "_Fact"), para tentar fazer uma aproximação do modelo conceptual com o modelo físico, onde se utilizam as "tabelas de factos".

Na figura 2.10 apresenta-se um exemplo das células constituintes de um facto com duas dimensões: a dimensão "Cliente", já apresentada, e a dimensão "Produto", que contém os níveis "Produto", "Categoria" e "Todos". Como têm de existir células para cada combinação de níveis das dimensões, temos 9 células neste facto, resultantes do produto cartesiano de todos os níveis das dimensões.

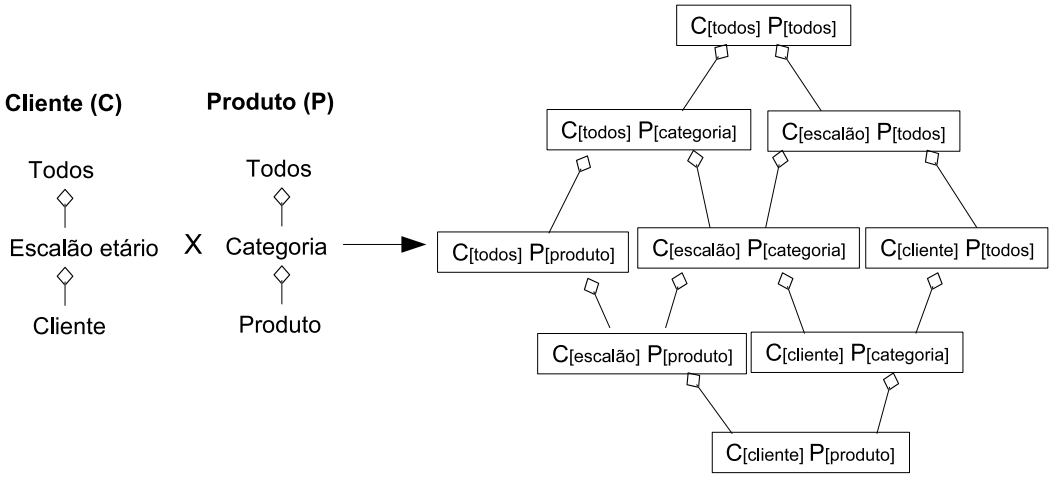

Figura 2.10: YAM^2: Exemplo de representação das células num facto com as dimensões "Cliente" e "Produto"

Podem ser definidas as relações entre os vários tipos de nós. O YAM^2 suporta 6 tipos de relações (incluindo a agregação, que já foi mostrada na figura anterior), cujas representações

são apresentadas na figura 2.11:

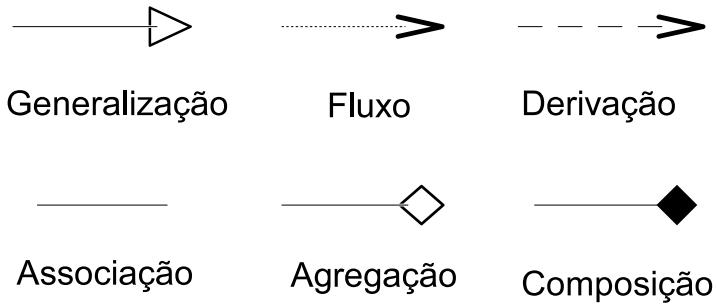

Figura 2.11: YAM: Representação das relações suportadas

- Generalização - relaciona dois elementos generalizáveis, um dos quais tem um significado mais específico que o outro.
- Fluxo - relaciona dois elementos do modelo, de tal forma que os dois representam diferentes versões da mesma coisa.
- Associação - define uma relação semântica entre dois elementos.
- Agregação - é um tipo mais forte de associação, em que um elemento representa parte do outro.
- Composição - se as partes não puderem ser partilhadas por diferentes todos, temos uma forma mais forte de agregação, conhecida por composição. Quando dois elementos estão relacionados por composição, isto implica que o elemento de origem faz parte do elemento de destino, de tal forma que se este último for destruído, os elementos que o compõem também serão destruídos.
- Dependência - esta relação contém vários subtipos, dos quais no YAM^2 só é utilizada a derivação. Esta representação ajuda a representar as relações entre elementos sob diferentes pontos de vista, ou seja, nas diferentes concepções do universo do discurso.

Na descrição do YAM são apresentadas todas as relações suportadas por cada par de elementos. Dado que aqui não se pretende reproduzir o estudo integral, faremos apenas um resumo de acordo com as relações que foram utilizadas no âmbito deste trabalho.

Se dois elementos podem ser relacionados através de determinada relação, será indicado se eles têm de pertencer ao mesmo elemento no nível conceptual superior (intra-relações) ou não (inter-relações). Dando um exemplo, a relação "generalização" entre dois níveis é uma inter-relação, pois estes níveis terão de pertencer a duas dimensões distintas (elemento do nível conceptual superior), para esta relação entre eles poder existir.

Para alguns nós existem relações que podem ser de ambos os tipos (intra ou inter), como por exemplo a associação entre duas dimensões. As relações utilizadas no âmbito deste trabalho foram as seguintes:

- Agregação - foi aplicada a níveis da mesma dimensão, tratando-se portanto de intra-relações (mas também pode ser aplicada a níveis de diferentes dimensões), e a factos, tratando-se de uma inter-relação, porque cada estrela tem apenas um único facto (os factos pertencem portanto a níveis conceptuais superiores - estrelas - diferentes).
- Generalização - foi aplicada a dimensões de diferentes estrelas (inter-relações), mas também pode ser aplicada a dimensões da mesma estrela (intra-relações).
- Derivação - aplicada a uma relação entre dimensão-dimensão e dimensão-facto dentro da mesma estrela (intra-relações).
- Associação - utilizada para relacionar dimensões e factos, e entre dimensões, dentro da mesma estrela (intra-relações).

Estão inerentes ao modelo conceptual as seguintes restrições de integridade:

- Superchave - corresponde ao conjunto de níveis numa célula
- Base - conjunto mínimo de níveis que são ”superchave´´ de uma célula, ou seja, o conjunto mínimo de níveis necessários e suficientes para identificar univocamente uma célula. Quando se usa mais dimensões do que as incluídas na base, pode-se obter espaços vazios na célula.
- Cubo - corresponde à associação de uma célula com uma base.

No âmbito deste trabalho são sempre apresentadas nos diagramas de nível intermédio as bases de cada facto. A restrição “Cubo” não será utilizada.

Existem também algumas condições necessárias para a sumarização da informação e que foi tida em conta para a concepção dos vários modelos apresentados:

1. Disjunção: Os subconjuntos de objectos a serem sumarizados têm de ser disjuntos.
2. Completude: A união dos subconjuntos devem constituir o conjunto inteiro.
3. Compatibilidade: O nível, o tipo de medida a ser sumarizado e a função estatística têm de ser compatíveis.

As condições 1 e 2 são dependentes das cardinalidades no total das relações entre dimensões, porque estas é que definem as categorias de agrupamento.

Alguns modelos proíbem relações M-N nas hierarquias de agregação, para evitar problemas na sumarização dos dados. Isto implica que uma parte (do nível menos agregado) só pode pertencer a um todo (do nível conceptual mais agregado). Mas, não há nenhum axioma mereológico[2] que proíba a partilha de partes entre vários todos. Por exemplo, poder-se-ia considerar que os vulgares ”kinder surpresa´´ poderiam pertencer a duas categorias de produtos: brinquedos e chocolates.

[2]Teoria ou estudo lógico-matemático das relações entre as partes e o todo e das relações entre as partes no interior de um todo.

Assim, no YAM^2 são permitidas hierarquias não estritas nas dimensões, e têm de ser levadas em conta para decidir a sumarização das medidas.

Outro problema nas cardinalidades tem a ver com as hierarquias ”esparsas´´, onde é permitido que para um mesmo nível existam várias relações partes-todo. Um exemplo possível será considerar o Mónaco simultaneamente uma cidade e um estado, sendo que neste caso podemos ”saltar´´ o nível ”estado´´, ou o nível ”cidade´´. Com esta situação surgem as hierarquias ”não cobertas´´, onde é possível saltar níveis conceptuais, ou as hierarquias ”non-onto´´, segundo o termo em inglês, onde é possível obter diferentes comprimentos dos caminhos desde a raiz até às folhas, dependendo das instâncias. Uma das soluções possíveis para este problema é a utilização de valores ”Dummy´´ para garantir o preenchimentos das partes.

Relativamente às restrições de sumarização o YAM^2 permite, nos diagramas do nível conceptual mais baixo, representar as condições de sumarização de uma maneira mais flexível do que apenas distinguir medidas ”não aditivas´´, ”semi-aditivas´´ e ”aditivas´´.

No entanto, para a descrição dos submodelos apresentados no nível conceptual mais baixo não foi utilizada neste trabalho a nomenclatura do YAM^2, pois considerou-se que era de difícil leitura para ser utilizada num modelo tão extenso.

Assim, optou-se pela utilização de estruturas tabelares, cuja informação foi definida com base nos diagramas do nível conceptual inferior do YAM, mas onde primeiro são apresentados os descritores de dimensões que são essenciais para a correcta compreensão das medidas e valores dos factos que são apresentados. Depois, são apresentadas as próprias medidas dos factos, indicando para cada uma a sua base e funções de agregação aplicáveis. Desta forma, fica completamente definido como é que cada uma das medidas consideradas nos factos podem ser agregadas.

Além da alteração efectuada aos diagramas de nível conceptual inferior, foi também definida uma alteração nos diagramas de nível conceptual intermédio que se apresentam na secção 4.4. A dimensão “Instalação” é apresentada num diagrama próprio, pois, dada a sua dimensão, estaria a sobrecarregar os diagramas dos sub-modelos se fosse sempre integralmente apresentada. Assim, em cada um dos submodelos esta dimensão será apresentada como uma ”caixa negra´´, tendo sido anteriormente descrita em detalhe, o que nos permite simplificar bastante o diagrama conceptual dos submodelos.

3

Trabalhos Relacionados - Sistemas de Informação Ambiental

Neste capítulo são apresentados alguns casos de estudo e trabalhos relacionados com os sistemas de informação ambiental.

O Ambiente é uma das temáticas mais complexas em termos da recolha de dados e da realização de análises que permitam ter uma visão completa, detalhada e extensiva sobre o seu estado actual. Quando se fala de Ambiente é necessário ter em conta que toda a actividade humana tem influência directa sobre o meio onde nos encontramos e que existem outros factores, muitas vezes invisíveis e esquecidos, que podem também influenciar o Ambiente como por exemplo o caso da actividade solar. Observando por exemplo a figura 3.1, que representa uma coisa tão comum como é o ciclo do carbono ([Obs]), verificamos que praticamente todas as componentes ambientais se encontram representadas:

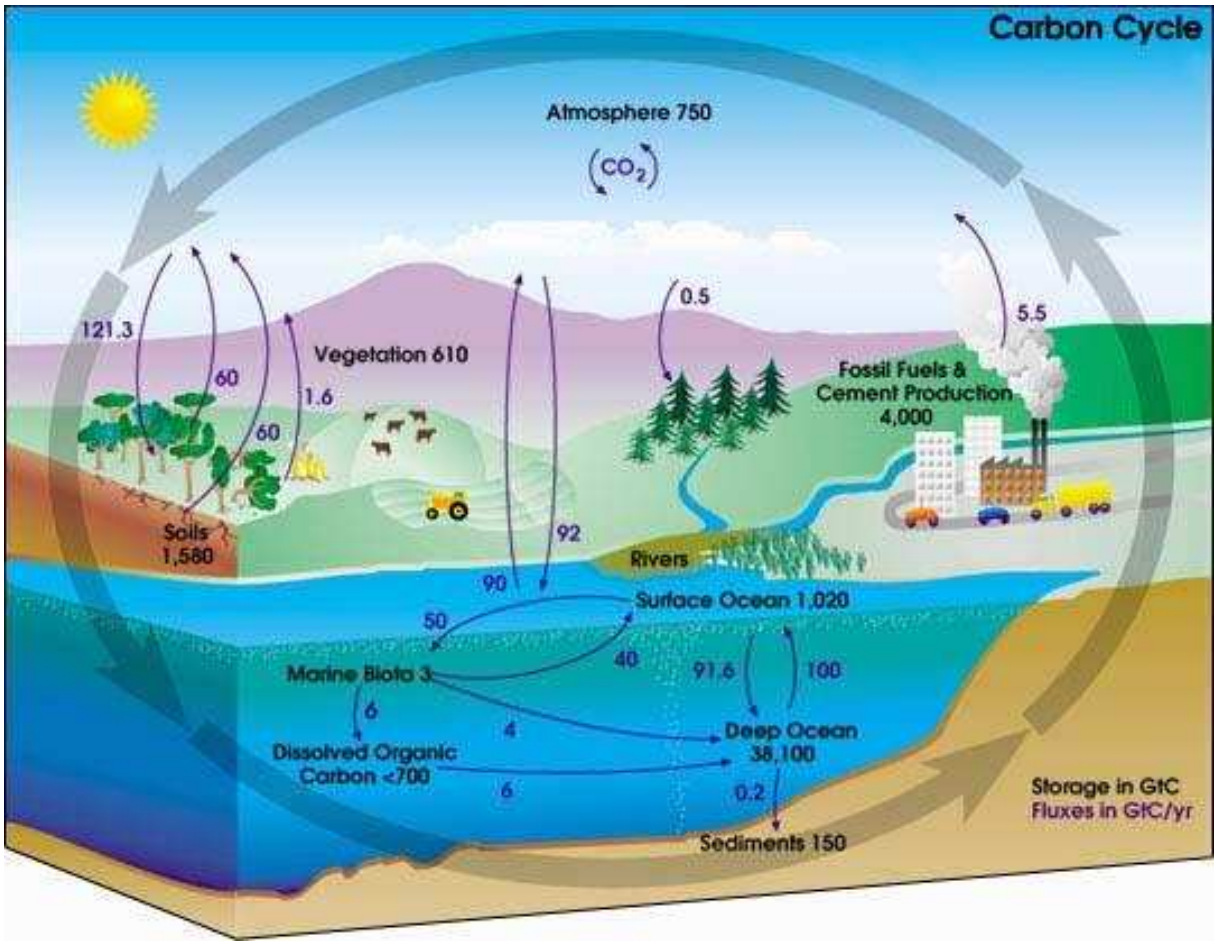

Figura 3.1: Diagrama do ciclo do carbono [SS]

- Atmosfera
- Solo
- Água
- Actividade humana
- Fauna
- Flora
- Sol

Ou seja, quando se está a falar de ambiente tem de se considerar tudo o que nos rodeia, tanto de origem natural como de origem humana, pois tudo o que acontece no planeta Terra tem de alguma forma impacto ao nível do Ambiente e mesmo alguns factores externos, como o Sol, têm de ser considerados.

No diagrama apresentado na figura 3.2 pretende-se esquematizar a interacção destes vários componentes.

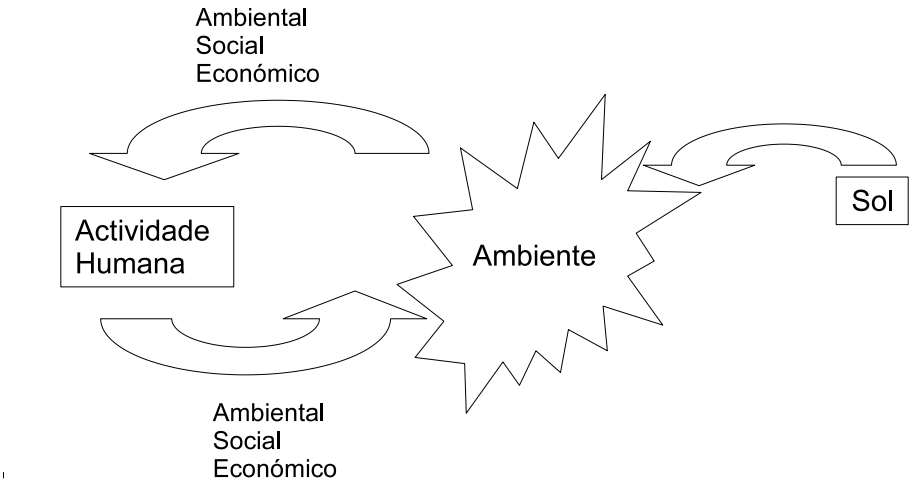

Figura 3.2: Interacção dos componentes que se relacionam com o Ambiente

Como se pode observar, o ambiente afecta tanto a actividade humana quanto esta tem impacto no ambiente. As indicações das palavras Ambiental, Social e Económico referem-se aos 3 eixos apresentados no início do capítulo 1, os eixos envolvidos no Desenvolvimento Sustentável.

No que diz respeito às actividades humanas, estas têm um papel muito importante em todos os sistemas naturais, e estão a provocar alterações ambientais ao nível local, regional e até mesmo global. Como exemplos de alterações provocadas a nível global temos por exemplo a alteração da atmosfera e do clima devido à desflorestação, queima de combustíveis fósseis e algumas actividades industriais. Ao nível regional temos como exemplo a bacia do rio Quequén Grande, um rio argentino cuja água, apesar da ainda baixa densidade de urbanização e industrialização que se verifica nesta zona, se tem vindo progressivamente a degradar devido ao desenvolvimento sócio-económico registado na zona nas últimas décadas [RCVT06].

A necessidade de pesquisa das contribuições e respostas humanas, muitas vezes chamadas "dimensões humanas", da mudança global motiva a investigação de várias questões relacionadas com este tema. As pesquisas sobre as dimensões humanas incluem estudos de motivadores tecnológicos, sociais, económicos e culturais da mudança global, cruzando muitas vezes áreas tão diversas como a composição atmosférica, o clima, os ciclos da água e do carbono, os ecossistemas, a utilização da terra e outras alterações globais.

Ou seja, o âmbito de análises possíveis sobre Ambiente pode ser tão vasto ou tão específico quanto se queira, e a quantidade de variáveis e informação que tem de ser incluída para fornecer resultados fiáveis pode ser muito extensa, o que torna a problemática do Ambiente muito complexa e o tratamento da informação relacionada assume as dimensões de um grande desafio. Além disso, o Ambiente é um importantíssimo factor a ter em conta para o alcance do desenvolvimento sustentável, concorrendo para a definição das políticas de desenvolvimento.

Assim, neste capítulo faz-se uma apresentação sobre matéria teórica de informação ambiental, incluindo uma breve incursão nos sistemas de apoio à decisão ambiental e, numa área mais específica, os sistemas de apoio à decisão espaciais. Foram seleccionados também alguns casos de estudo relacionados com esta temática para serem apresentados, devido à sua relevância, e no final efectua-se um pequeno sumário a ressaltar as conclusões que foram consideradas a partir destes casos de estudo.

3.1 Indicadores e relatórios de estado do ambiente

A disponibilização de informação ambiental é um dever das autoridades públicas [REA05], com o objectivo de fomentar a consciencialização dos cidadãos e a sua participação na resolução dos problemas do ambiente. A produção de relatórios periódicos sobre o estado do ambiente é uma prática que se tem vindo a generalizar e é um modo de concretizar este princípio. O estado do ambiente é normalmente traduzido através de um conjunto de indicadores, que abrangem várias áreas de estudo e que estão a ser adoptados ou adaptados por vários países, para a produção de Relatórios de Estado do Ambiente (REA).

3.1.1 Indicadores do ambiente

Os indicadores ambientais são medidas do estado do ambiente. O seu objectivo é fomentar a consciencialização do ambiente e mostrar progressos na direcção do desenvolvimento sustentável [Agea].

Os indicadores adoptados pela *Environment Agency*, cujo objectivo é apresentado na frase anterior, combinam dados transversais a um conjunto de medidas ambientais e foram inferidos a partir de várias fontes, incluindo governamentais. Os indicadores ambientais têm três funções básicas: simplificação, quantificação e comunicação. Os utilizadores podem escolher indicadores do conjunto disponível para preencher as suas necessidades. Existem alguns critérios para a selecção de indicadores [UA97], nomeadamente:

1. Com base no seus objectivos:
    - Providenciar uma base para comparações internacionais.
    - Serem simples e facilmente interpretáveis.

- Mostrar as tendências ao longo do tempo.
- Ser sensíveis às alterações que pretendem medir, principalmente as relacionadas com a actividade humana.
- Ter um alvo ou um limite com o qual possam ser comparados, para que os utilizadores possam compreender os valores a que estão associados.
- Ser facilmente medíveis, ou seja, de obtenção relativamente simples.

2. Com base na sua profundidade analítica, um indicador deve ser:
   - Bem fundamentado teoricamente em termos técnicos e científicos.
   - Baseado em *standards* internacionais e ter um consenso internacional quanto à sua validade.
   - Propício à associação com modelos económicos, de previsão e de sistemas de informação.
3. Com base na sua capacidade de serem medidos, os dados requeridos para suportar o indicador devem:
   - Estar disponíveis rápida e facilmente.
   - Ser adequadamente documentados e com uma qualidade conhecida.
   - Ser passíveis de sofrerem actualizações regulares.

No caso específico do REA português são considerados 3 tipos de indicadores [REA05]:

- descritivos, que reflectem o que está a acontecer ao Ambiente e ao Homem;
- de desempenho, que efectuam comparações com metas nacionais e internacionais;
- de eficiência, que relacionam diversos elementos da cadeia causal.

Para alguns dos indicadores incluem-se comparações internacionais (*benchmarking*), que ilustram a situação nacional no contexto da União Europeia (UE) ou dos países da Organização para a Cooperação e Desenvolvimento Económico (OCDE).

3.1.2 Relatório do estado do ambiente

O objectivo da realização de relatórios sobre o estado do ambiente é providenciar informação exacta, actualizada e acessível sobre as condições ambientais, suas tendências e pressões para as áreas a que dizem respeito [Gov05].

Neste contexto, a disponibilização de informação sobre diversas matérias por parte das autoridades públicas é um dever, que facilita a consciencialização dos cidadãos e estimula a sua participação como principais aliados na resolução dos problemas, entre os quais os do ambiente.

A produção de relatórios periódicos sobre o estado do ambiente é uma prática que se tem vindo a generalizar na maioria dos países, constituindo um modo de concretizar os referidos princípios. Mesmo algumas organizações como por exemplo a ONU têm, através do seu Programa Ambiental das Nações Unidas (UNEP na sigla original), publicado relatórios anuais e abrangentes dedicados ao ambiente, além de várias outras publicações relativas a determinados países ou zonas, também dedicadas a este tema [Prob].

Também algumas entidades privadas se têm dedicado à realização de relatórios periódicos sobre o estado do ambiente, como por exemplo a BC Hydro, a maior instalação para produção de electricidade na Columbia Britânica (servindo 94% da população desta província) [Hyd06].

No território nacional temos alguns exemplos de empresas que publicam relatórios de estado do ambiente, como é o caso da Solvay Portugal [S.A], uma empresa de fabrico e/ou comercialização de produtos químicos de base foi uma das empresas portuguesas que aderiu aos princípios do programa voluntário Actuação Responsável, uma iniciativa para melhorar o desempenho da indústria química em matéria de saúde, segurança e ambiente, lançada pela Associação Portuguesa de Empresas Químicas (APEQ) em 1993.

Ao nível das autoridades portuguesas, desde 1987 têm vindo a ser editados relatórios anuais do estado do ambiente pelo IA [dA04]. A informação que é incluída nos REA varia conforme o seu âmbito e objectivo, mas por exemplo na Noruega foram consideradas as seguintes áreas de informação para serem incluídas no REA [UA97]:

- Alterações climáticas
- Qualidade do desenvolvimento urbano
- Degradação da camada de ozono
- Biodiversidade
- Acidificação - A chuva ácida é causada pelo enxofre proveniente das impurezas da queima dos combustíveis fósseis e pelo nitrogénio do ar, que se combinam com o oxigénio para formar dióxido de enxofre e óxido de nitrogénio. Na atmosfera, estes componentes reagem com o vapor de água e formam ácido sulfúrico e ácido nítrico, caindo depois para a terra na forma de chuva ou neve, ou como depósitos secos de sais. O dióxido de enxofre e o óxido de nitrogénio existem naturalmente na atmosfera, mas a industrialização provocou um grande aumento do seu nível.
- Eutrofização - fenómeno causado pelo excesso de nutrientes num recurso aquífero, o que leva à proliferação excessiva de algas. Estas, ao entrarem em decomposição, provocam o aumento do número de microorganismos e a consequente deterioração da qualidade da água. Pode abranger rios, lagos, baías, estuários, etc. [Wikb] As principais fontes de eutrofização são as actividades humanas industriais, domésticas e agrícolas. Por exem-

plo, os fertilizantes usados nas plantações, ao dissolverem-se e infiltrarem-se nas águas subterrâneas, são arrastados até aos cursos de água mencionados.

- Paisagens naturais e culturais.
- Resíduos.
- Contaminação tóxica.
- Recursos florestais.
- Recursos aquáticos vivos (peixes, crustáceos e moluscos).

Existem ainda alguns aspectos ambientais que estão directamente relacionados com o desenvolvimento sustentável [EEA02]:

1. Emissões de gases de estufa.
2. Geração de electricidade com base em recursos renováveis.
3. Volume de transporte (toneladas e passageiros por quilómetro) relativo ao Produto Interno Bruto (PIB).
4. Tipos de transporte utilizados (incluindo não motorizados) dentro da necessidade total de transporte.
5. Qualidade do ar urbano.
6. Resíduos municipais.
7. Eficiência energética.

Todos estes aspectos têm de ser considerados na tomada de decisão ambiental, conjugando as questões ambientais, económicas e sócio-culturais.

3.1.3 Recolha de informação sobre o Estado do Ambiente

Apesar da informação ambiental ao nível de cada país ter limitações, ainda são principalmente os governos nacionais que detêm o poder para recolher dados e definir políticas relativas a problemas ambientais [GT95].

No entanto, existem várias outras fontes de dados ambientais, como por exemplo o Sistema de Monitorização Ambiental Global do UNEP (UNEP/GEMS na sigla original) [Prob], a Organização da Comida e Agricultura das Nações Unidas (FAO na sigla original) [FAO], a Divisão de Estatística das Nações Unidas (UNSD na sigla original) [Nat06], a Agência Internacional da Energia (IEA na sigla original) [OEC06] e o Centro de Monitorização para a Conservação Mundial (WCMC) do UNEP [UW].

Quanto às formas de obtenção da informação ambiental na sua origem, ou seja, recolha e registo dos valores e dados ambientais necessários, estas são variadas e diversificadas. Grande

parte da informação é obtida directamente pelos institutos e entidades governamentais responsáveis pela obtenção de informação. Um bom exemplo são os dados demográficos, que em Portugal são da responsabilidade do Instituto Nacional de Estatística (INE).

Outro conjunto de informação é obtida directamente a partir das instalações industriais que, devido a compromissos devidamente legislados, têm periodicamente de entregar dados relativos aos poluentes emitidos. Um bom exemplo desta situação é o European Pollutant Emission Register (EPER), que pretende estabelecer um registo europeu de dados comparáveis relativos às emissões de poluentes provenientes dos vários países, e que em Portugal é da responsabilidade do IA.

Existem também recolhas directas de amostras ambientais (por exemplo, para testar solos) efectuadas por organizações e entidades devidamente credenciadas para o efeito.

Por último, referem-se as designadas redes de monitorização, muito utilizadas para avaliação da qualidade do ar [REA05]. Os critérios mínimos de monitorização em Portugal encontram-se devidamente regulamentados, o que permitiu a delimitação de Zonas e Aglomerações, sendo as Aglomerações áreas de maior concentração populacional e dispondo de pelo menos duas estações - uma urbana de tráfego e outra urbana de fundo. Nas Zonas, e independentemente das concentrações populacionais observadas, há pelo menos uma estação para avaliar a poluição de fundo e a poluição causada por eventos naturais.

Os dados recolhidos por estas vias serão depois utilizados na criação de sistemas de informação ambiental, servindo de suporte à realização de análises de dados ambientais.

3.2 Sistemas de Apoio à Decisão Ambiental (SADA)

Uma classe particular dos sistemas de suporte à decisão são os Sistemas de Apoio à Decisão Ambiental (SADA), que pretendem apoiar as decisões ambientais. Os problemas ambientais têm uma complexidade própria, pois integram muitas variáveis distintas, quer de natureza quantitativa, quer relativas a aspectos políticos, legislativos e económicos. À medida que a protecção do ambiente e o desenvolvimento sustentável vão ganhando prioridade na maioria dos países actualmente, os SADA vão tendo maior relevância pelas administrações públicas e pela comunidade científica.

Os objectivos e características de um SADA têm em conta alguns aspectos complementares aos de um SAD convencional, relacionados com problemas ambientais. Estas são características relacionadas com aspectos naturais, como as características espaciais, temporais, periodicidade, e também características relacionadas com a própria tomada de decisão (informação abrangente a várias áreas, responsabilidades distribuídas na resolução de problemas, multiplicidade de critérios, etc).

As actividades relacionadas com o ambiente podem, de uma forma geral, ser agrupadas em cinco áreas funcionais [Bat]:

- previsões e análise de risco;
- planeamento a médio e longo prazo;
- monitorização e vigilância;
- gestão de crises e emergências;
- avaliação de danos e reclamações.

É geralmente reconhecido que os SADA são aplicações complexas, muitas vezes integrando diferentes tecnologias e requerendo esforços intensos de pesquisa e desenvolvimento.

Uma das características principais que têm de preencher, enquanto SAD, é a integração. Portanto, têm de ser capazes de abranger várias fontes de informação ou bases de dados, várias representações de problemas, ou modelos, e têm de ter uma interface orientada para a resolução de problemas.

Ao nível dos dados e informação base, os SADA têm de integrar muita informação variada, que se revela por vezes ser incompatível, proveniente de várias fontes.

Também têm de conseguir lidar com pequenas questões técnicas, tais como permitir diferentes unidades de medida, lidar com faltas de documentação e permitir rastreio de ficheiros ou arquivos associados.

O processo de extracção de dados é extremamente complexo [MV01], pois nestes sistemas os dados residem frequentemente em fontes externas, e as instituições designadas para tomar decisões ambientais estratégicas (Ministério do Ambiente, Secretarias de Estado do Ambiente, etc.) não são responsáveis pela recolha dos dados, que são normalmente efectuadas por institutos, universidades ou organizações ambientais.

3.2.1 Sistemas de Apoio à Decisão Espaciais(SADE)

Um Sistema de Informação Geográfica (SIG) é um sistema informático capaz de capturar, armazenar, analisar e apresentar informação geograficamente referenciada [USG].

Com um SIG pode-se associar informação a dados de localização e podem-se criar camadas de informação que, ao serem combinadas, permitem perceber o funcionamento das várias componentes em conjunto [tgtgis]. Assim, cria-se a capacidade de relacionar diferente informação num contexto espacial, de forma a responder às perguntas colocadas e alcançar conclusões sobre estas relações, bem como modelar cenários para testar determinadas hipóteses [Giv].

Actualmente a utilização de SIG pode ser combinada com outros tipos de dados, como por exemplo dados metereológicos [Szn03], estatísticas criminais, dados demográficos, etc.,

para responder a um conjunto muito abrangente de questões, relacionadas com diversas áreas (negócios, segurança, saúde pública, ...).

Temos como exemplo de utilizações do SIG temáticas tão abrangentes como: a melhoria das rotas de transporte das frotas, a previsão de carga energética necessária em determinadas zonas, análises precisas sobre o potencial de dano de um furacão, melhoria da segurança pública através da monitorização das luzes de segurança e identificação de zonas problemáticas para optimizar a distribuição de recursos policiais [Cou].

Um sistema SIG tem, na sua base, os seguintes componentes essenciais [Lou]:

- *Hardware* - suporta as várias actividades de um SIG, desde a recolha dos dados até à sua análise. O ponto central do equipamento é a estação de trabalho onde o *software SIG* é executado e onde pode ser ligado algum equipamento auxiliar que seja necessário. Com a divulgação do SIG através da *Web*, os servidores *Web* também se tornaram uma parte importante do equipamento de um SIG.
- Um SIG precisa de diferentes pacotes de *software*:
  - Aplicação SIG - é essencial para a criação, edição e análise de dados espaciais e de atributos, contendo uma grande variedade de funções SIG para este efeito.
  - Componentes SIG - permite a criação de aplicações de *software* com um objectivo específico, mais limitadas em termos das suas capacidades analíticas.
  - Utilitários - permitem a execução de algumas actividades auxiliares à utilização do SIG, como por exemplo converter ficheiros para formatos mais facilmente utilizáveis.
- Dados - a captura de dados é normalmente a parte mais consumidora de tempo na utilização da tecnologia SIG. Os dados podem provir de fontes comerciais, educacionais, governamentais, pesquisa própria, etc. A condição da superfície, atmosfera e da camada abaixo da superficie da Terra podem ser examinadas através da introdução de dados provenientes de satélites num SIG. Desta forma, é possível observar as variações que decorrem na Terra ao longo de dias, meses ou anos.
- Pessoas - um conjunto de pessoas treinadas em análise espacial que saibam utilizar *software* SIG são essenciais para o sucesso de um sistema SIG.

Muitas áreas das aplicações de suporte à decisão dizem respeito a dados geográficos. Assim, as técnicas SIG começam a ter impacto nas aplicações de suporte à decisão [Kee04]. Enquanto as aplicações SIG podem conter a informação necessária para apoiar a decisão, estes são normalmente sistemas com um âmbito de utilização genérico e não focados numa decisão particular, além de necessitarem de pessoas com conhecimentos na sua utilização.

Assim, torna-se por vezes necessária a criação de modelos direccionados para determinados

problemas para suportar a tomada de decisão, por vezes através de processamento adicional e de integração com dados não espaciais, constituindo aplicações mais fáceis de utilizar e mais acessíveis. Desta forma, os SIG são utilizados na construção dos SAD.

Os SADE são portanto um subconjunto importante dos SAD, com potencial para um crescimento rápido, devido aos desenvolvimentos tecnológicos e à disponibilidade de tecnologia de baixo custo para a manipulação de dados espaciais. Os problemas que se pretendem analisar com um SADE precisam de ter uma componente espacial, relacionada com pelo menos um dos aspectos da tomada de decisão, embora possam também incorporar dados não espaciais.

Um exemplo da utilização de um modelo SADE é a selecção do local para uma fábrica. Alguns dos critérios envolvidos nesta decisão terão uma natureza espacial, como por exemplo, a localização das escolas (para as famílias dos trabalhadores) e potenciais zonas de descarga dos resíduos. As operações espaciais do SIG podem ser utilizadas para fornecer uma lista de lugares adequados, mas o responsável pela decisão poderá ter de ponderar outros factores, pelo que pode ser necessário utilizar técnicas para tomada de decisão envolvendo vários critérios. Ou seja, utiliza-se um modelo tradicional para quantificar as várias alternativas e as operações espaciais para identificar o impacto da decisão em termos espaciais.

3.3 Casos de estudo

Foram encontrados vários casos de estudo relacionados com a área da informação ambiental. No entanto, optou-se neste documento por apresentar com detalhe apenas aqueles que, de alguma forma, estavam directamente relacionados com a temática deste trabalho ou que, devido à complexidade e interesse das problemáticas que abordavam, pareceram ser manifestamente relevantes, tendo merecido portanto uma análise mais detalhada.

É de salientar também que os casos de estudo apresentados neste documento estão ou em fase de desenvolvimento ou já com protótipos em funcionamento ou mesmo em produção, pelo que são bons candidatos a serem contabilizados como referências interessantes.

Além do mais, revelou-se difícil encontrar projectos ou trabalhos relacionados com esta área que estivessem documentados com exaustividade, podendo portanto ser analisados ao pormenor para se compreender completamente as problemáticas apresentadas e encontrar as associações existentes para o trabalho apresentado neste documento. Entre aqueles que não foram incorporados neste documento referimos, por exemplo, o WATERSHEDSS [Gro], o Decision Support System for Air Operating Permits [oEQT] e o ONTOWEDSS [Cec01].

3.3.1 Le Select

Em 2001 foi realizado um artigo ([MV01]) onde se descreve o desenvolvimento de um trabalho no contexto do projecto Ecobase [pm], iniciado em 1999, que propõe a combinação de tecnologias de mediação e DW, como suporte para criar um SADA.

A tecnologia de mediação usa o paradigma de publicação de dados, ou seja, a capacidade para disponibilizar os dados na *Web*, providenciando acesso uniforme aos utilizadores e aplicações ambientais, independentemente do formato de armazenamento dos dados (ficheiros de texto, folhas de cálculo, tabelas relacionais, etc.). A mediação combina algumas tecnologias actuais para encontrar, transformar e disponibilizar dados, de uma forma que o sistema pode sempre evoluir e crescer, através da inserção de novas fontes de informação.

O DW trabalha com dados históricos e replicados, necessários para as análises requeridas na tomada de decisão nas organizações.

3.3.1.1 Arquitectura do mediador Le Select

O Le Select é um protótipo de *framework* para acesso a dados de naturezas heterogéneas e para invocar programas de processamento dos dados em ambientes Intranet/Internet.

O Le Select tem uma arquitectura completamente distribuída, onde se podem distinguir duas entidades principais: os *sites* de publicação e os clientes. Os dados e programas podem ser publicados em *sites*, desde que exista um servidor Le Select em execução, sendo que o significado de publicação aqui é tornar os dados disponíveis. Aqueles que publicam dados são designados publicadores e os que lhes acedem são chamados clientes.

Os utilizadores e as aplicações vêem os dados publicados como tuplos em tabelas relacionais, cujas fontes de dados podem não ser bases de dados: ficheiros de texto, *spreadsheets*, etc.

A informação sobre como aceder e como localizar os dados está disponível numa espécie de invólucro, cujo código e definição dos dados é da responsabilidade do publicador. A principal tarefa deste invólucro é colocar os dados em tabelas relacionais que serão utilizadas pelo Le Select, sendo constituído por um conjunto de classes de JAVA, com os ficheiros de definição dos dados em XML. Os invólucros podem ser escritos de forma genérica para poderem ser reaproveitados e existem também invólucros para aceder aos programas.

O Le Select oferece diferentes mecanismos de acessos aos dados: drivers JDBC, protocolos FTP e HTTP. Os clientes ou aplicações que desejem utilizar dados ou programas em ambiente Internet/Intranet precisam de se ligar aos servidores correspondentes, usando a componente cliente do Le Select, responsável por efectuar a ligação entre o utilizador e o *site* de publicação. Os utilizadores podem também utilizar *browsers Web* para explorar os dados e programas pub-

licados e é possível utilizar SQL para procurar nas tabelas exportadas pelos vários invólucros distribuídos, numa só pesquisa.

Os programas podem estar numa máquina e os dados serem processados noutra, caso em que o sistema envia dados para o *site* onde está o programa, coordena a sua execução e envia de volta os resultados ao cliente, sob a forma de tabelas relacionais.

O Le Select não tem um repositório central de dados nem um esquema global; em vez disso, existem vários servidores que cooperam para fornecer acesso aos dados e programas, providenciando interoperabilidade entre as fontes de dados distribuídas, heterogéneas e autónomas num ambiente Internet/Intranet.

3.3.1.2 Aplicações ambientais

O paradigma de publicação de dados oferecido pelos tecnologia de mediação tem permitido aos fornecedores de dados ambientais a disseminação de dados para um grande número de utilizadores, que os podem visualizar através de *browsers*, independentemente do seu formato ou localização.

Como os dados ambientais apresentam um elevado nível de heterogeneidade e são armazenados numa variedade de repositórios distribuídos por todo o mundo, é facilmente dedutível que as aplicações ambientais podem ser baseadas numa arquitectura distribuída, cujos componentes podem ser integrados de forma incremental. Esta arquitectura tem de ter em contas aspectos como

- sistemas - diferenças entre sistemas operacionais e *hardware*;
- sintaxe - diferenças nas representações dos dados;
- estrutura - que suporta os diferentes formatos e organizações dos dados;
- semântica - o que diz respeito à interpretação dos dados, onde as diferenças no significado dos dados são dependentes do vocabulário e terminologias usadas para expressar a interpretação dos conteúdos da informação e as suas relações.

O *middleware* Le Select consegue gerir os três primeiros níveis mencionados, portanto pode ser considerado uma ferramenta adequada para construir diferentes aplicações ambientais.

Além de providenciar a publicação de dados e a sua posterior visualização através de um *browser*, o Le Select suporta a criação de aplicações cliente, o que motivou o desenvolvimento de uma aplicação para extrair dados para outro repositório, encorajando as organizações a usar dados produzidos por outras organizações.

Neste contexto, esta ferramenta deve ser capaz de extrair dados de diferentes fontes, transformá-los e carregá-los num repositório homogéneo. Mesmo que replicados, estes dados podem ser extremamente úteis para os gestores tomarem as suas decisões ambientais.

3.3.1.3 Arquitectura proposta

Uma vez publicados, os dados e metadados ficarão disponíveis para exploração através de um *browser*, a partir do qual os desenhadores do Data Warehouse Ambiental (DWA) podem decidir quais os atributos de facto importantes para incluir no DW. A figura 3.3 apresenta uma arquitectura geral do sistema proposto, que inclui as seguintes componentes:

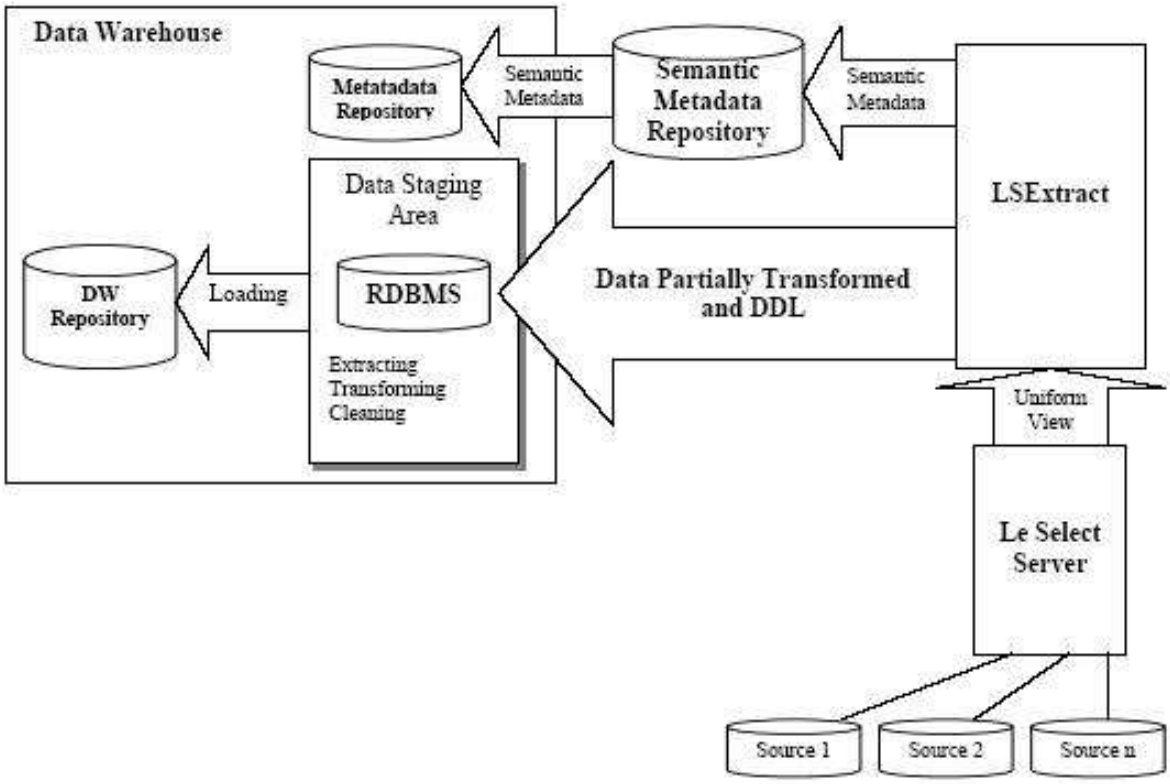

Figura 3.3: Arquitectura geral do Le Select

- Fontes - representam fontes de dados autónomonas, heterogéneas e distribuídas, tal como ficheiros de texto, folhas de cálculo, tabelas relacionais, etc.
- Le Select - traduz as fontes de dados, oferecendo às aplicações clientes um acesso uniforme aos dados publicados.
- LSEstract - fornece a extracção de dados e metadados das fontes traduzidas pelo Le Select. Executa algumas transformações, inserindo dados numa tabela relacional (uma área de *staging*) e os metadados semânticos num repositório de metadados do DW. Esta ferramenta captura os metadados estruturais das fontes traduzidas para criar uma Data Definition Language (DDL) que é usada depois para criar o esquema de dados da área de *staging*.
- Repositório de metadados semânticos - é constituído por várias directorias de ficheiros, suportando os seguintes formatos: texto, HTML ou XML.

- Data Warehouse - contém a área de *staging* e os repositórios de dados e metadados. Após os dados serem carregados na área de *staging* é possível usar ferramentas de ETL para realizar o carregamento final para o DW. Depois, os metadados do repositório de metadados semânticos podem também ser copiados para o repositório de metadados do DW.

3.3.2 Projecto SIMAGE

Este projecto [MNL] é financiado pelo Ministério do Ambiente italiano (orçamento aprovado em 2001) e tenciona melhorar a gestão de risco nas áreas industriais de Itália. Os principais objectivos deste projecto são:

- A criação de redes de comunicações de qualidade do ar, harmonizadas, para as áreas industriais de Brindise e Taranto (Itália), incluindo a integração e optimização das redes já existentes, a instalação das novas estruturas e instrumentos de monitorização da poluição do ar, definição de procedimentos de controlo de qualidade e os laboratórios de controlo associados.
- O estabelecimento e desenho de um centro de coordenação nacional para troca de informação ambiental no que diz respeito ao ar, água e qualidade do solo, interligado com as maiores áreas de risco de Itália, em particular, as áreas de Brindisi, Taranto, Porto Marghera, Priolo-Augusta, Gela, Milazzo.
- O desenvolvimento de um sistema piloto para monitorização e controlo do transporte de substâncias perigosas principalmente por estrada, incluindo um exercício de avaliação de rastreio e tecnologias de comunicação móvel, e a implementação de sistemas locais em Brindisi, Taranto e Porto Marghera para controlo de tráfego e gestão de emergências.

Para explorar a informação disponível, e fortalecer a acção do Ministério do Ambiente, o Instituto para Protecção e Segurança do Cidadão (IPSC) irá também fornecer a este Ministério um Sistema Integrado de Ferramentas Data Warehouse (SIFDW).

3.3.2.1 Desenho do Sistema Integrado de Ferramentas Data Warehouse (SIFDW)

Uma vez que o projecto pretende avaliar as necessidades de desenho de ferramentas de suporte, a abordagem que tem sido adoptada respeita um nível estratégico que considera o risco provocado pela indústria como parte do problema relacionado com um esquema regional complexo.

Este nível estratégico requer uma visão detalhada do efeito, ao longo do tempo e espaço, das políticas que estão a ser definidas em termos de impacto ambiental e consequências sócio-económicas. Dentro do contexto caracterizado por múltiplos objectivos, e conflituosos, é particularmente interessante aplicar a MCDM (ver secção 2.1).

Verificou-se, no entanto, que a abordagem típica de desenhar o sistema de MCDM baseado numa base de dados estática e num sistema computacional pode apresentar algumas limitações. Neste caso específico, agravadas pelo facto de que os dados acumulados pelas várias organizações e companhias não se mostram fiáveis o suficiente para suportar o processo de tomada de decisão.

Devido à heterogeneidade e multitude de repositórios de informação distribuídos requeridos por um processo de tomada de decisão tão complexo, a MCDM deve ser baseada numa estrutura distribuída cuja informação pode ser acedida dinamicamente, de acordo com os requisitos do problema. E, a avaliação dos vários cenários possíveis requer uma interacção entre os critérios e alternativas, enquanto os gestores obtêm uma melhor visão sobre as suas preferências e expandem ainda mais os seu conjunto de alternativas.

Por esta razão, os SAD têm de ser construídos de uma forma que permitam alterações rápidas e fáceis, e tem de ser seguido um processo iterativo que envolva alterações contínuas, pelo que se propõe melhorar a arquitectura do SAD usando um sistema baseado num DW.

Considerando essa estrutura, pode-se assumir que o nível inferior do DW está directamente relacionado com os dados ambientais fonte. O processo de agregação ou extracção pode ser executado definindo procedimentos de pesquisas ad-hoc, ou técnicas de *Data Mining*.

Outra forma de agregação pode ser a implementação de modelos que retirem os seus *inputs* de diferentes bases de dados operacionais e armazenem os seus *outputs* num repositório de dados centralizado. Consequentemente, só o nível de topo está relacionado com o processo de apoio à decisão.

A qualidade dos dados pode ser considerada na perspectiva do ciclo de vida da informação, pois a partir de um pedido de informação os dados não tratados podem ser recolhidos e transformados, através do DW, em informação especificamente moldada para os utilizadores finais. Só considerando as expectativas dos utilizadores e a relevância da informação obtida se pode analisar se o sistema responde aos pedidos dos utilizadores finais. Assim, o ciclo de vida de análise torna-se importante na optimização do SIFDW baseado em DW.

Actualmente o sistema protótipo do SIFDW está em fase de desenvolvimento. A ferramenta será baseada em metodologia *Multicriteria analysis* (MCA), especificamente moldada para auxiliar os gestores a lidar com a sustentabilidade das áreas industriais.

Na figura 3.4 apresenta-se o esquema conceptual do protótipo, onde é possível verificar que se pretende que o SIFDW vá além dos SAD típicos, porque não se trata simplesmente da integração de um tipo de modelo que auxilia na síntese de informação. As necessidades dos utilizadores não são aqui conhecidas à partida, e os requisitos para a definição de uma política industrial podem ser diferentes de acordo com os objectivos dos responsáveis pela tomada de decisão e podem variar ao longo do tempo.

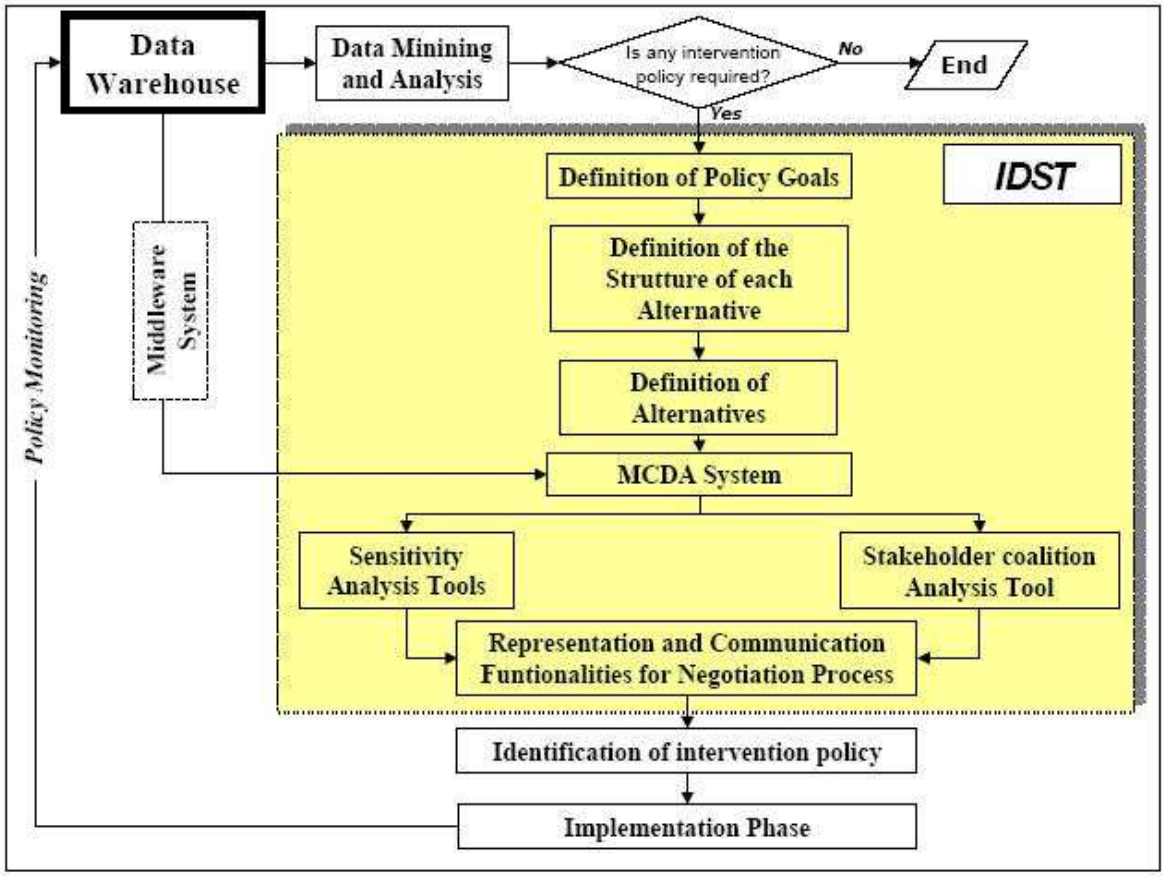

Figura 3.4: Esquema conceptual do protótipo SIMAGE

Por este motivo, existe o ciclo para definir os objectivos da política, e as posteriores ferramentas de análise (de sensibilidade e dos responsáveis pela tomada de decisão), antes de se considerar identificada a política e se passar à fase de implementação para incorporação no DW. Para validar a abordagem e características do protótipo SIFDW, foi usada uma abordagem participatória que envolve substitutos de gestores, industriais e criadores de políticas.

De acordo com o objectivo principal do SIMAGE foi organizado um fórum (FARI, 2001) a nível internacional, intra-disciplinar e inter-disciplinar entre cientistas, industriais, criadores de políticas e Organizações Não Governamentais (ONG) sobre alternativas para o desenvolvimento industrial sustentado em Itália e, consequentemente na Europa.

Os especialistas concordam que normalmente os criadores de políticas e as autoridades apreendem o ambiente não globalmente, mas apenas em termos dos seus principais componentes (exemplo: ar, água, solo) e, consequentemente, a gestão ambiental reflecte esta visão desintegrada. Assim, a definição de um SIFDW baseado em sistemas DW pode ser uma abordagem com sucesso, mesmo que existam vários problemas ainda a resolver.

3.3.3 Instalação Pantex

A instalação Pantex[1] [SKL] pertence ao United States Department of Energy/National Nuclear Security Administration, estando actualmente a ser gerida pela BWX Technologies, Inc [BT].

Várias áreas da instalação estão submetidas a investigações da Resource Conservation and Recovery Act (RCRA), porque algumas operações efectuadas na instalação na altura da II Guerra Mundial provocaram a contaminação do ambiente em vários sítios da instalação (a Pantex tem uma área de 40,5 Km^2, dos quais 23,9 Km^2 são considerados zonas de segurança). Além disso, as licenças de resíduos da instalação contêm alguns requisitos de monitorização e prestação de dados.

Os Environmental Remediation Services (ERS) e Regulatory Compliance Departments (RCD) da instalação, responsáveis pela investigação, limpeza e encerramento dos sítios problemáticos, têm de gerir um conjunto de informação ambiental sempre crescente, que requer uma aproximação sistemática e apresentação de dados analíticos, mapas, registos, fotografias.

A informação inclui resultados analíticos de amostra para vários meios ambientais, relatórios de restauração e suporte ambiental, mapas, desenhos de instalações, registos de poços e furos, e fotografias dos sítios que estão a ser ou já foram geridos pela Pantex.

Os gestores de projecto da ERS e os cientistas requerem acesso a informação diária proveniente de quase 7000 localizações, incluindo 150 veios de água subterrâneos, poços de gás, praias fluviais, furos no solo, unidades de gestão de resíduos sólidos e outros sítios de amostra ambiental. Para acomodar esta necessidade foi criado um *Web site* interno para providenciar aos gestores e cientistas acesso aos dados, que suportam a monitorização trimestral e relatórios de dados e para providenciar dados da instalação para simulações e cálculo de risco para a RCRA.

3.3.3.1 Componentes do sistema

O Data Warehouse Ambiental é constituído por um conjunto de elementos ligados (ver figura 3.5). Os componentes incluem uma Base de Dados Ambiental Integrada (BDAI), o DW analítico, uma base de dados geográfica, um *site* interno e um conjunto de pesquisas e ficheiros *batch* para actualizar os dados.

- Base de Dados Ambiental Integrada (BDAI) - os dados analíticos das amostras ambientais recolhidas na Pantex são armazenados nesta base de dados, desenvolvida em 1996. Os dados armazenados incluem amostras químicas do solo, veios aquáticos subterrâneos, dados geotécnicos recolhidos de amostras de subsuperficie e medições do nível da água

[1] A instalação Pantex tem 5 objectivos operacionais principais: Montagem e desmontagem de armas nucleares, avaliação de armas, pesquisa e desenvolvimento de produtos altamente explosivos que rodeiam os componentes nucleares das armas e local de armazenamento temporário de plutónio [Pan]

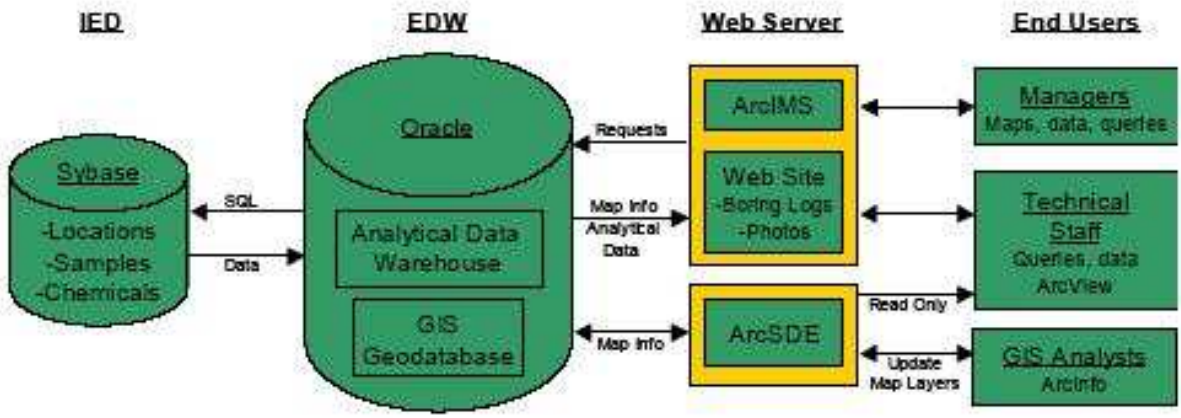

Figura 3.5: Arquitectura dos componentes do DWA da instalação Pantex

recolhidas em poços, num total de quase 2 milhões de registos.

Esta base de dados foi desenvolvida pela Sybase em 1996, e representa a segunda iteração de uma base de dados ambiental para a instalação Pantex. São adicionados novos dados diariamente, quer a partir de entregas de dados dos laboratórios contratados, ou carregando dados recebidos por contratantes. O pessoal ERS verifica continuamente os dados armazenados na base de dados para garantir a sua qualidade e, sempre que necessário, são efectuadas correcções.

- DW analítico - é uma base de dados Oracle®separada que armazena dados retirados das tabelas da BDAI.

  Foi criado para simplificar as muitas tabelas da BDAI num repositório *standard*, tanto para pesquisas como para efeitos de SIG. Adicionalmente, ficaram disponíveis outras fontes de dados além da BDAI, que podem ser integradas no DW analítico sem fazer alterações a nenhuma das aplicações existentes. Os dados armazenados no DW analítico foram verificados através de um conjunto de regras, para assegurar a validade e consistência dos dados.

  Os resultados são ligados numa estação principal, num ficheiro que permite a representação espacial dos dados. Como a BDAI é actualizada diariamente com novos registos e correcções, o DW analítico tem de ser actualizado regularmente.

- Base de dados geográfica (BDG) - foram integrados no DW analítico conjuntos de dados do SIG ambiental da Pantex, utilizando o *software* ESRI ArcSDE, que permite a aplicações ArcGIS e ArcIMS armazenarem, gerirem e acederem a dados espaciais directamente na base de dados Oracle®.

  A base de dados geográfica ArcSDE providencia algumas características benéficas, incluindo a gestão centralizada dos dados SIG, versões de níveis, multi-acessos a partir do

ArcGIS e do ArcIMS, ligação dinâmica aos dados do DW analítico e rastreio dos dados armazenados.

As primeiras 3 características permitem aos funcionários manter toda a informação SIG numa única localização, fazendo a actualização da base de dados geográfica ArcSDE a partir dos seus postos de trabalho, garantindo que têm as características mais actualizadas dos mapas.

Além disso, o pessoal da TI efectua cópias de apoio e mantém a base de dados, permitindo ao pessoal do SIG focar-se na manutenção dos dados. Como a base de dados geográfica é centralizada, pode ser partilhada com outras organizações na instalação, que começam a tirar partido das aplicações SIG.

As ligações dinâmicas ao DW analítico permitem que os resultados de amostras para veios aquáticos subterrâneos, solo, águas de superfície e águas residuais sejam convertidas em características de pontos ArcSDE a partir das coordenadas da BDAI. Quando uma localização de amostra é pesquisada numa aplicação SIG, são apresentados todos os resultados de amostras existentes para aquela localização.

- *Web site* interno - permite aos funcionários técnicos e de gestão interagir com a informação no DW analítico, requerindo apenas um *browser* para poder ser utilizado. A página principal providencia ligações para as 5 funções disponíveis actualmente: pesquisa de dados analíticos, visualizador SIG, registos de poços e furos, fotografias e representações gráficas de dados temporais.

3.3.4 Envirofacts

O Envirofacts é a principal base de dados de acesso público da United States Environmental Protection Agency (EPA) [(EP], e reúne subconjuntos de 7 dos 13 sistemas nacionais de dados contemplados no ”Reinventing Environmental Information (REI) Action Plan´´, e inclui informação sobre água potável, poluição do ar, derrames tóxicos e lixo perigoso [oS99].

A ideia subjacente à criação desta base de dados é colocar informação ambiental de qualidade disponível na *Internet*, para todos a poderem ver. Desta forma, foi possível fomentar a melhoria da qualidade dos dados fornecidos, pois optou-se por publicar todos os dados, independentemente da sua condição ou veracidade, o que teve como resultado que os responsáveis pela recolha de dados passaram a ter uma preocupação efectiva sobre a sua autenticidade e qualidade, e também porque todas as discrepâncias encontradas pelos utilizadores são relatadas para que se proceda à sua correcção [(EP00].

Assim, o Envirofacts encontra-se disponível através da *Web* [Ageb], permitindo aos utilizadores a consulta rápida e fácil de informação sobre actividades ambientais que afectem o

ar, a água e o terreno em qualquer ponto dos Estados Unidos da América.

Através do Envirofacts é possível verificar quais as instalações na vizinhança que emitem poluentes ou que lidam com materiais perigosos, onde se localizam os sítios altamente poluídos, bem como o seu estado de limpeza. Para criar este *site* a EPA seguiu um conjunto de regras:

- Todos os dados chegam por via electrónica, sendo recebidos e tratados a nível central.
- Os dados são armazenados em DW integrados.
- Os metadados são também capturados e disponibilizados aos utilizadores.
- Existem formulários para que possam ser enviadas observações para facilitar o relato das discrepâncias.
- Os dados devem seguir determinadas normas.
- Foram disponibilizadas linhas de atendimento para ser feito o relato de dados instantâneo.
- Os motores de pesquisa são os mais actuais possíveis.
- Após os dados estarem no DW, passam a ser propriedade do estado, uma regra necessária porque, caso contrário, não poderiam ser legalmente corrigidos os erros de tipografia e outros erros óbvios.

O Envirofacts tem também uma interface SIG de acesso fácil para os utilizadores com menos conhecimentos técnicos, que lhes permite efectuar pesquisas e visualizar a informação num mapa, através da utilização do código ZIP, cidade, estado, *county*.

Tem também disponível uma lista de tópicos que permitem oferecer informação mais detalhada, como por exemplo água, resíduos tóxicos, ar, radiação, terra.

Adicionalmente, tem também uma interface para os utilizadores com maiores conhecimentos técnicos ao nível do SQL, que permite, caso o registo do utilizador tenha sido efectuado e aceite, a execução de pesquisas directas sobre as bases de dados que alimentam o Envirofacts, através da *Internet*. Esta interface permite também a selecção de dados para construção de relatórios tabulares, ou o *download* de informação, através de ficheiros CSV[2].

Este *site* recebia, em 2000, cerca de 1,6 milhões de *hits* por mês, tendo-se tornado o repositório de dados para muitas organizações, muitas vezes ultrapassando os próprios sistemas internos [(EP00].

[2]Comma Separated Values

3.4 Nota finais

O Le Select, tendo sido apresentado no âmbito de um projecto de investigação Franco-Brasileira na área de sistemas de informações, o Ecobase [Rel01], apresenta uma solução interessante para a problemática da recolha e integração dos dados, permitindo a sua publicação através de formatos normalizados e facilmente acessíveis. No entanto, não fazia parte do âmbito deste projecto a proposta de utilização de uma *framework*, apesar de interessante, e a utilização de uma ferramenta deste tipo revela-se de difícil aplicabilidade, sendo necessária a alteração de factores culturais, ou a utilização de instrumentos políticos para promover a utilização de um instrumento tão transversal.

O projecto SIMAGE aborda as problemática da obtenção de dados ambientais a partir da fonte, através de redes de monitorização e da integração da informação num só repositório.

O caso de estudo da instalação PANTEX apresenta com detalhe a arquitectura de um DW que se encontra em exploração e que reúne dados de 7000 localizações diferentes, incluindo também componentes de SADE.

O Envirofacts permite perceber a problemática da integração de informação proveniente de vários estados, aplicado ao caso específicos dos Estados Unidos da América, mas permitindo extrapolar e perceber algumas das dificuldades sentidas na integração de dados ao nível europeu.

4

Modelo conceptual

Neste capítulo apresenta-se o modelo conceptual resultante do estudo efectuado no âmbito desta dissertação.

Os modelos conceptuais dão ao utilizador muito mais informação sobre a realidade que está a ser modelada, e aproximam-se mais da sua forma de pensar, o que é especialmente importante nas actividades de processamento analítico, devido à natureza imprevisível das pesquisas desejadas pelo utilizador nestes ambientes. Este tipo de utilizadores não pode ficar restrito a um conjunto limitado de pesquisas predefinidas. Na realidade, eles têm de conseguir criar novas pesquisas específicas, a maioria das vezes baseando-se apenas nos metadados. Assim, é essencial que um modelo conceptual mostre tanta informação semântica quanto possível.

Neste capítulo serão apresentados os conceitos básicos relacionados com Ambiente utilizados nos modelos propostos, detalha-se o âmbito do trabalho apresentado nesta dissertação e os requisitos de modelização que foram identificados, bem como os próprios modelos que corporizam o estudo apresentado nesta dissertação.

4.1 Conceitos do domínio a modelizar

Antes de se passar à apresentação dos modelos conceptuais, é conveniente debruçarmo-nos um pouco sobre os conceitos e definições utilizados nesta dissertação, e que contribuem para explicar a modelização efectuada.

Na figura 4.1 apresentam-se, de forma simplificada, os vários componentes de uma actividade industrial.

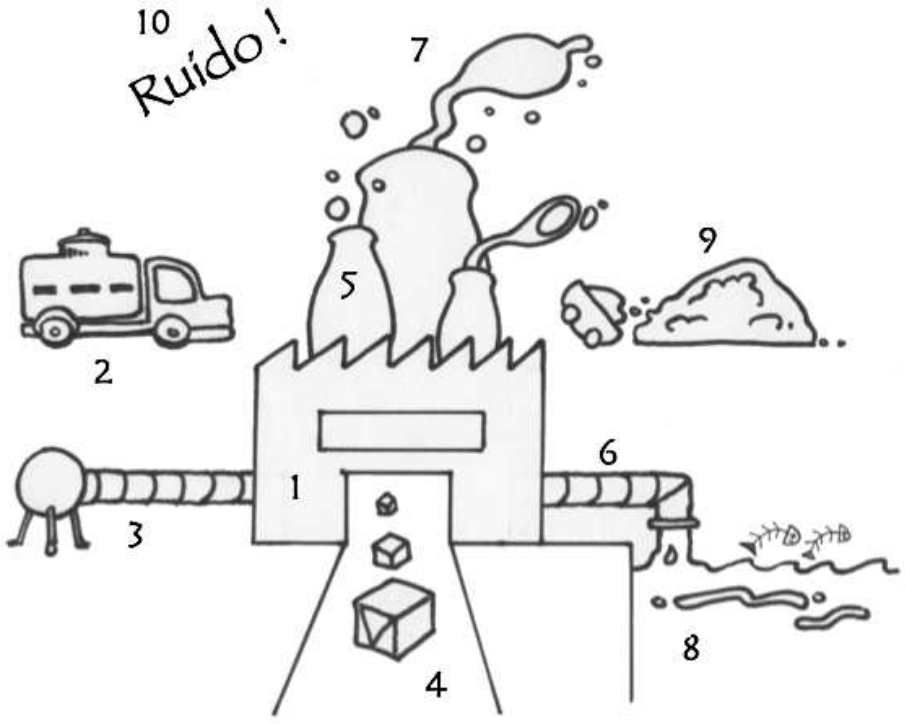

Figura 4.1: Componentes da actividade industrial

Nesta figura a instalação encontra-se assinalada com o número 1 e representa uma instalação de índole industrial, mas que pode na realidade ser de outra índole (agrícola, serviços) desde que seja relevante sob o ponto de vista ambiental. Eis alguns exemplos de instalações de diversas actividades: *(i)* central termoeléctrica; *(ii)* fábrica de papel; *(iii)* instalação agropecuária; *(iv)* estação de tratamento de resíduos perigosos ou de águas residuais; *(v)* lavandaria (lavagem a seco); *(vi)* oficina de pintura de automóveis.

Sob o ponto de vista ambiental importa conhecer para cada uma destas instalações a sua localização geográfica (preferencialmente de forma o mais precisa possível), e quais as actividades económicas nela desenvolvidas.

Os elementos da figura 4.1 assinalados pelos números 2, 3 e 4 representam a visão económica da instalação, a qual necessita para a sua operação de matérias primas (2), energia e água (3). O resultado da actividade da instalação sob o ponto de vista económico está ilustrado pela saída de produtos da instalação e assinalado como 4.

As matérias primas podem eventualmente ter um particular interesse sob o ponto de vista ambiental, pois podem ser potencialmente nocivas para o ambiente, como é o caso dos compostos orgânicos voláteis (COV), por exemplo presentes nas matérias primas usadas pelas lavandarias e oficinas de pintura de automóveis. Por outro lado, quer as matérias primas, quer os produtos finais ou intermédios podem merecer uma particular atenção do ponto de vista da sua perigosidade, quer em termos de armazenagem, quer em termos da sua manipulação ou transporte (SEVESO).

Os restantes elementos assinalados na figura 4.1 ilustram alguns dos aspectos relevantes sob o ponto de vista de gestão ambiental: para as chaminés (assinaladas com o número 5), em geral designadas como fontes pontuais de emissão atmosférica, interessa conhecer vários aspectos como a altura, o diâmetro, o caudal médio, o tipo de filtros e os equipamentos que contribuem para cada fonte; de modo similar interessa conhecer as características das descargas para o meio liquido (rios, lagos, etc) que se encontram assinaladas com o número 6.

Sob o ponto de vista ambiental é já hoje essencial controlar as emissões específicas de certos poluentes, quer para o ar quer para meio líquido (assinalados respectivamente com os números 7 e 8), para manter registos nacionais e internacionais de quantidades emitidas de uma lista de poluentes referenciados.

Os números 9 e 10 assinalam mais duas preocupações sob o ponto de vista ambiental: resíduos sólidos e, de um modo geral, a contaminação dos solos; e ruído, particularmente importante em meios urbanos.

Importa referir que esta visão simplificada, embora útil sob o ponto de vista de comunicação e síntese, deixa de fora algo de essencial. Uma dada actividade económica recorre a uma série de actividades industrais (actualmente existe uma tabela classificativa destas actividades) que

podem ser executadas recorrendo a diferentes processos industriais (igualmente tabelados e de acordo com as actividades industriais).

A instalação é caracterizada por um conjunto de Actividades Económicas (CAE), das quais uma delas é considerada a principal. Cada instalação tem uma ou mais actividades industriais, que caracterizam o tipo de produção efectuada.

Uma instalação necessita de matérias primas, água e energia para trabalhar. Se os solventes fizerem parte do conjunto de matérias primas necessárias, essa instalação terá actividades incluídas no conjunto de actividades que emitem Compostos Orgânicos Voláteis (COV).

Se essa instalação cobrir os requisitos definidos pela legislação da Prevenção e Controlo Integrado da Poluição (PCIP) - detalhada mais à frente - ela terá um conjunto de actividades incluídas na lista de actividades PCIP. Para produzir os seus objectivos, as instalações utilizam um conjunto de processos industriais (designados por processos Nose-P no caso do PCIP), em que para cada processo existe uma capacidade instalada da instalação.

Dando um exemplo concreto, para a actividade industrial pertencente à lista PCIP que inclui as "Instalações de tratamento de superfície de metais e matérias plásticas que utilizem um processo electrolítico ou químico, quando o volume das cubas utilizadas nos banhos de tratamento realizado for superior a 30 m^3 (2.6)", existem instalações que utilizam os processos industriais (neste caso, Nose-P) "Desengorduramento de metais sem utilização de solventes (105.01.01)", "Electrodeposição (105.01.03)", "Decapagem (105.01.06)" e "Outros tratamentos de superfície em metais (105.01.20)", enquanto outras utilizam os processos "Desengorduramento de metais sem utilização de solventes (105.01.01)", "Electrodeposição (105.01.03)" e "Fosfatação (105.01.05)".

Uma das componentes que se pretende avaliar do ponto de vista da análise de dados é exactamente as diferenças do ponto de vista ambiental (e, muitas vezes, económico) entre os diferentes processos utilizados.

Estas entidades e as suas relações são apresentados no grafo conceptual da figura 4.2.

Este grafo também apresenta, embora sob uma forma esquemática, o subconjunto dos componentes que foram mencionados na descrição da actividade industrial, e que são considerados no âmbito da modelação.

Como resultado desses processos, são produzidas emissões de determinados poluentes para um determinado meio (por exemplo, ar, água), e que podem ser medidas através dos métodos de medição (por exemplo, cálculo, estimativa), e é produzido o produto final que pode ser medido através do volume de produção, e que está associado ao cálculo da capacidade efectivada da instalação.

Para explicar melhor as capacidades efectivada e instalada da instalação, podemos dizer que a capacidade instalada reflecte a capacidade máxima da instalação para realizar produção,

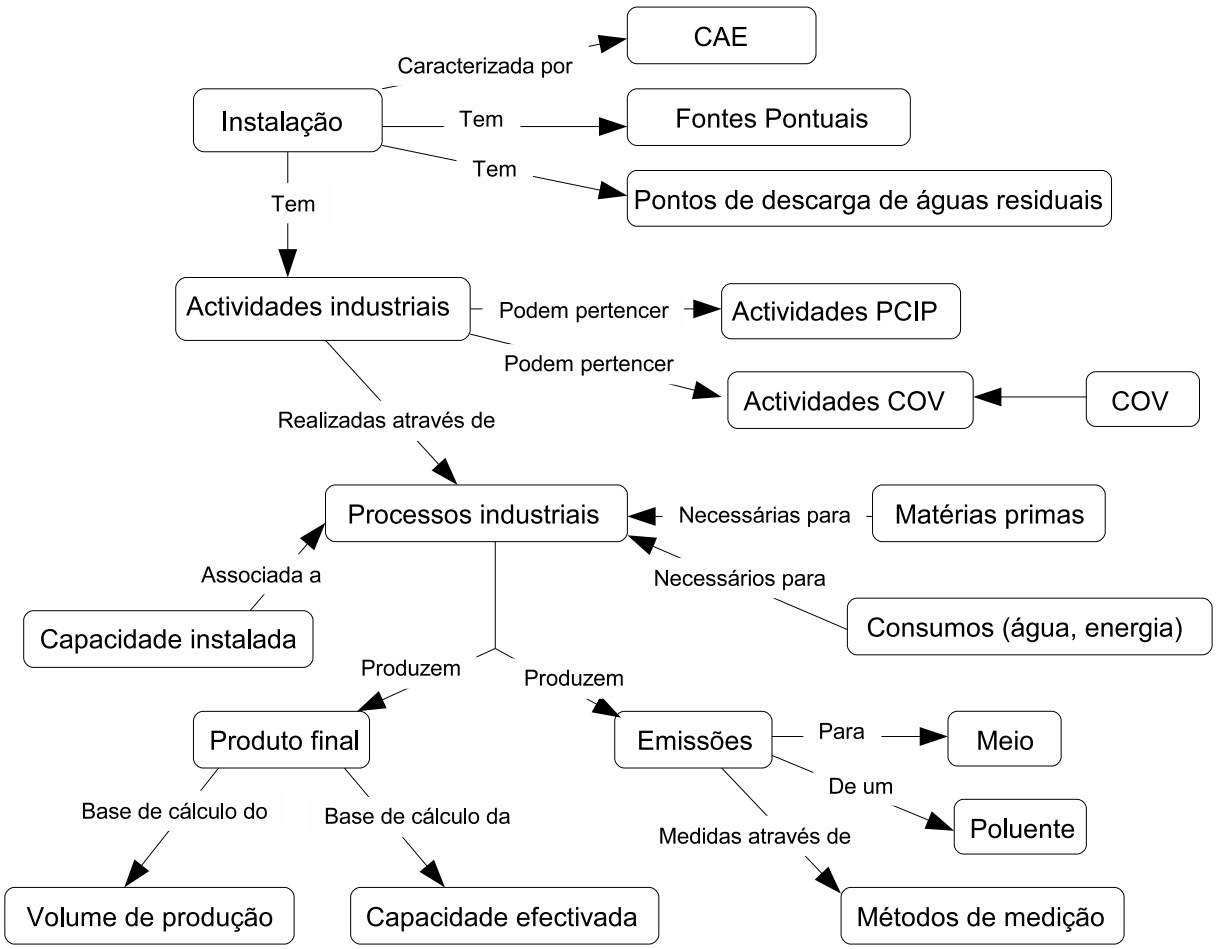

Figura 4.2: Grafo de conceitos associados à instalação

enquanto a capacidade efectivada pode ser vista como a percentagem da capacidade da instalação que é efectivamente utilizada. A capacidade instalada é indicada pelos operadores da instalação e está relacionada com a própria infraestrutura da mesma, enquanto a capacidade efectivada está relacionada com a produção final realizada, e a percentagem da capacidade da instalação que é utilizada para obter essa produção final (caracterizada pelo volume de produção).

Relativamente a este grafo de conceitos (figura 4.2), é de referir que ele foi desenhado tendo em conta o modelo ideal, mas não necessariamente o actualmente implementado.

Por exemplo, actualmente os consumos de água e energia (que passarão a ser designados por dados de funcionamento) não são recolhidos ao nível do processo industrial, mas sim ao nível da instalação. Naturalmente que esta diferença entre o nível da recolha destes dados de funcionamento (ao nível da instalação) e os dados de produção (volume de produção, capacidade instalada e capacidade efectivada, ao nível dos processos) impede a comparação entre ambos ao nível de processo, o que limita a análise comparativa da eficiência dos vários processos em termos de consumos/resultados.

Como as emissões também são recolhidas ao nível do processo, podem ser comparadas para

cada processo com os dados de produção, mas não com os dados de funcionamento, limitando mais uma vez a análise comparativa entre processos.

Na próxima secção passaremos a explicar o universo de dados abrangidos pelos sub-modelos propostos (dos quais os conceitos gerais acabaram de ser explicados) e as motivações para as opções tomadas.

4.2 Âmbito

Conforme já foi referido anteriormente, o Ambiente tem duas componentes principais:

- Efeito da actividade humana sobre o ambiente
- Estado do ambiente

Também foi referido que neste trabalho seria abordada apenas a componente relativa ao efeito da actividade humana sobre o ambiente. Existem vários motivos para esta ter sido a componente escolhida:

- Os dados relativos à actividade humana são essenciais, embora não suficientes, para se conseguir obter a informação sobre o estado do ambiente.
- A avaliação do impacto da actividade humana no ambiente é fundamental para se controlar a actividade industrial (e outras). As próprias normas internacionais exigem que este tipo de avaliação seja realizada, quer para se efectuar licenciamento das indústrias (temos como exemplo o licenciamento PCIP), quer para se obterem inventários das emissões realizadas (caso do EPER, por exemplo), que por sua vez também vão fornecer novos dados para a continuação da avaliação dos efeitos sobre o ambiente. E, através da contínua comparação dos efeitos da actividade humana sobre o ambiente quando se utilizam processos industriais diferentes, é possível fomentar a adopção das melhores técnicas e processos disponíveis, possibilitando muitas vezes obter vantagens também ao nível económico, além do ambiental.
- Por outro lado, a própria recolha de dados se encontra mais disponível para a vertente de avaliação do impacto da actividade humana no ambiente do que para a vertente de estado do ambiente porque, para o primeiro caso, os próprios operadores responsáveis pelas instalações têm de fornecer dados (como acontece para o EPER), mas no segundo caso muitos dos dados são obtidos através da realização de estudos, mais ou menos específicos e através da exploração das redes de monitorização do ambiente.

No IA existem vários mecanismos de recolha de dados para avaliação do impacto da actividade humana no ambiente, dos quais foram seleccionados conjuntos ou subconjuntos de informação para serem modelizados e incluídos nesta dissertação.

Conforme já referido na secção 3.1.3, uma das funções do IA é a implementação do EPER (European Pollutant Emission Register) em Portugal. O EPER pretende analisar ao nível europeu quais os níveis de emissões registadas pelas instalações industriais mais poluidoras de cada país para determinados poluentes, ou seja, pretende analisar os maiores valores de emissões ao nível dos países europeus para o conjunto de poluentes definido.

Para se efectuar a selecção das emissões pretendidas foi definido pela UE um conjunto de limiares para os poluentes abrangidos pelo EPER (um total de 50 poluentes, 37 dos quais aplicáveis ao ar, 26 aplicáveis a água). O objectivo destes limiares, segundo o '"Guidance Document for EPER Implementation" [fE00], é que não sejam reportados valores insignificantes de poluição e permitir, ao mesmo tempo, que pelo menos 90% do total de emissões industriais da Europa seja reportado.

Os limiares destes poluentes foram definidos com base nos valores industriais relativos ao Reino Unido (Inglaterra e País de Gales), Alemanha e Países Baixos, de forma independente dos processos industriais (NOSE-P).

Fazem também parte dos dados recolhidos no EPER (fornecidos pelos operadores das instalações) os valores dos volumes de produção e capacidades das instalações abrangidas, bem como volumes energéticos consumidos. Estes dados foram também considerados no âmbito deste trabalho, corporizando os submodelos dos dados de produção e dados de funcionamento, que serão apresentados ao longo deste documento.

O IA é ainda o responsável [dA04] pela Prevenção e Controlo Integrado da Poluição (PCIP), que impõe o licenciamento ambiental das entidades com actividades poluidoras, que sejam abrangidas pelo Anexo I do Decreto-Lei número 194/2000 de 21 de Agosto. Em Portugal, o licenciamento PCIP foi corporizado através da entrega de um formulário preenchido pelos operadores das instalações ambientais e da entrega de documentos técnicos.

O formulário PCIP abrange praticamente todos os agentes de poluição: emissões para o ar, água e solo, poluição sonora, resíduos, descargas de águas residuais e fontes pontuais, corporizando um conjunto de informação extensa e de tratamento complexo. Por esse motivo, não foi incluído nesta dissertação o formulário completo, mas sim um subconjunto de informação, seleccionada de acordo com a opinião e experiência dos técnicos do IA, tendo os critérios de selecção tido em conta a transversalidade ao nível das instalações incluídas no conjunto de dados, e a complexidade da informação abrangida.

Uma outra área de actividade do IA está relacionada com os Compostos Orgânicos Voláteis (COV), abrangendo resíduos gerados por indústrias que utilizam como matérias primas produtos nocivos, como solventes orgânicos.

Relacionada com os COV apenas será considerada a ficha de recensamento das entidades, já que os restantes formulários ou não estão sequer definidos, ou não estão estabilizados, pelo

que não constituem uma fonte credível de análise para o estudo apresentado nesta dissertação.

4.3 Factores de modelação

Quando foi iniciada a análise da informação disponível no IA para ser efectuada a proposta dos submodelos para os vários conjuntos de dados, foram definidas algumas condições que deveriam ser preenchidas pelo modelo proposto, para garantir que este fosse abrangente e extensível. Assim, a modelação conceptual que se vai propor na próxima secção deverá ter em conta estas condições, nomeadamente:

- Garantir a transversalidade das dimensões (eixos de análise) propostas, para que seja possível o cruzamento da informação proveniente de vários sectores do IA.
- Propor uma definição clara da dimensão tempo, que mais uma vez seja uma dimensão transversal, permita facilmente realizar análises de dados para diferentes períodos e acomodar as indisponibilidade de dados para alguns períodos. Por exemplo, para o EPER ainda não estão a ser recolhidos dados para todos os anos, pois a periodicidade desta recolha de informação ainda não é anual (prevê-se que em 2007 talvez o seja). Existem períodos, como por exemplo o ano 2003, para os quais não existem dados. Isto tem particular importância quando se pretender efectuar análises evolutivas ao longo do tempo. Também tem de ser previsto que existem dados no formulário EPER que não são obrigatórios pelo que podem existir alguns dados, mas não todos os dados para um mesmo período de tempo.
- Ser facilmente expansível para suportar diferentes tipos de informação adicional, como por exemplo informação de capacidades e volumes de produção e alguma informação sócio demográfica que é relevante para a realização de análises abrangentes.
- Acomodar alteração das tabelas classificativas, pois algumas listas classificativas já sofreram alterações ao longo do tempo (por exemplo, a lista de CAE). O DW tem de conseguir apresentar a informação correcta ao longo do tempo, com as listas utilizadas na data a que se reportam os dados, ou, caso essa seja a opção dos utilizadores e se houver conversão entre as diferentes listas, com a versão actual da lista após conversão.
- Suportar as dimensões lentamente alteráveis ao longo do tempo (SCD). É um caso diferente do anterior, já que não se prende apenas com listas de valores, mas vários atributos de uma dimensão podem sofrer alterações, mesmo não sendo listas classificativas.
- Sendo a informação considerada neste modelo de carácter ambiental, deverá ter uma vertente geográfica muito forte, para que as análises possam ser referenciadas geograficamente da forma mais precisa possível. Embora a actual componente geográfica da

informação ainda não seja muito forte (apenas as instalações são geograficamente referenciadas), foram já previstas possibilidades de extensão e enriquecimento da componente geográfica dos dados para que, caso esta seja a opção escolhida pelo IA no futuro, o modelo já acomode esta característica.

- Ter em conta os diferentes níveis de agregação dos dados (por exemplo, processos NOSE-P, 5 dígitos versus 7 dígitos).
- Suportar o rastreio dos dados. Neste momento as infraestruturas operacionais do IA ainda não suportam esquemas de rastreio da informação, mas deverá de ser prevista ao nível do DW a possibilidade de implantação de um esquema deste tipo no futuro.

A abordagem de análise e modelação multidimensional adoptada foi a *bottom-up*, defendida por Kimball [1ke]. O processo seguido para a definição do modelo conceptual partiu da análise da informação disponível relativa aos vários processos do IA, e definição dos *data marts* associados a cada área de informação (ou "negócio") do IA. O processo de inclusão de toda a informação num DW único não está ainda abrangido por este trabalho, conforme foi já referido na secção 1.2.

4.4 Sub-modelos considerados

Para a apresentação deste modelo na sua forma conceptual foi utilizada a nomenclatura do YAM^2, embora com as modificações que foram apresentadas na secção 2.6.

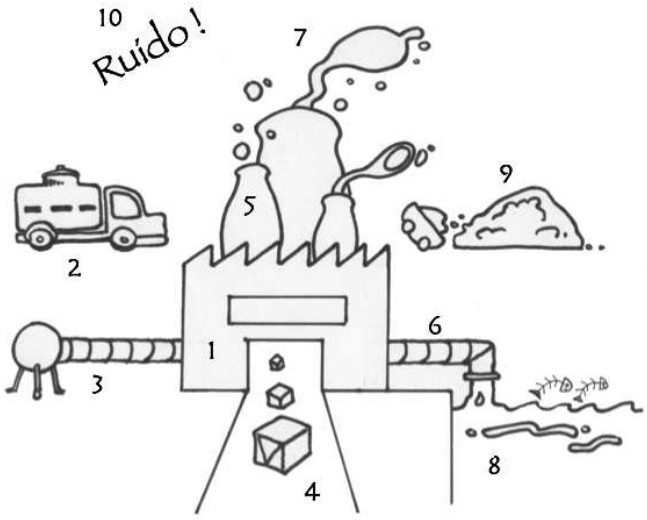

Figura 4.3: Componentes da actividade industrial

O mapeamento entre a figura 4.1, aqui novamente reproduzida, e 4.2 apresentadas na secção 4.1, permite elencar as zonas de informação que foram consideradas no âmbito deste trabalho, e que corporizam os vários sub-modelos propostos.

As áreas 7 e 8 referem-se às **emissões de poluentes da instalação** (assinalada como área 1). No âmbito deste trabalho são consideradas as emissões para os meios "ar" e "água".

A área 4 apresenta os **dados de produção**, o resultado da actividade industrial da instalação.

As áreas 2 e 3 da figura 4.3 dizem respeito aos **dados de funcionamento**, um dos submodelos considerados. Através da recolha de dados possibilitada pelos formulários EPER e PCIP é possível analisar consumos de água e energia ao nível da instalação.

A área 5 apresenta as **fontes pontuais** associadas à instalação e dados relativos aos **equipamentos que contribuem para as fontes pontuais**, mais um sub-modelo.

A área 6 da figura 4.3 está relacionada com as **descargas de águas residuais**.

A área 9 diz respeito aos **resíduos**, uma área que não foi incorporada neste trabalho. Embora exista alguma informação relativa a resíduos no formulário PCIP, dada a extensão do formulário, foi necessário seleccionar partes a serem incluídas nesta tese e esta não foi uma das áreas seleccionadas pelos técnicos do IA como tendo maior prioridade para incorporação neste trabalho. A mesma situação se verifica para o ruído (área 10).

Além dos 5 sub-modelos já referidos (emissões de poluentes, dados de produção, dados de funcionamento, descargas de águas residuais, fontes pontuais e equipamentos que contribuem para estas fontes pontuais) foi também considerada informação relativa a **dados demográficos**, e os dados relativos aos **COV**. Os COV estão implícitos nas áreas 2, 3 e 9 da figura 4.3, pois estão relacionados com os resíduos gerados por consumo de solventes, ou seja, quando uma instalação utiliza solventes como parte das matérias primas.

A apresentação de cada uma das estrelas constituintes dos sub-modelos será feita de uma forma *top-down*, começando por serem explicados os diagramas YAM^2 de nível conceptual mais alto, seguindo-se os de nível intermédio e em último lugar o nível conceptual mais baixo (com maior detalhe).

O primeiro sub-modelo a ser apresentado será o das emissões de poluentes, devido à sua abrangência. É também um dos sub-modelos mais complexos, dado o número de dimensões que utiliza (nove), e o número de tabelas de factos que contém (duas).

Para poupar espaço nos diagramas de nível conceptual intermédio, e também porque se trata de uma dimensão extremamente complexa, e que portanto justifica uma descrição mais detalhada, a dimensão **Instalação** será explicada em separado, no decorrer da apresentação do sub-modelo de emissões. Nos restantes sub-modelos, no nível conceptual intermédio esta dimensão aparecerá apenas sob a forma de uma caixa sombreada, não se apresentando novamente o seu detalhe.

Embora sejam utilizados nos diagramas os sufixos "_Dim" e "_Fact" para diferenciar as dimensões e os factos (conforme referido na secção 2.6), para facilitar a leitura do texto descritivo dos sub-modelos estes sufixos não serão utilizados, quando for explícito qual dos conceitos

está a ser referido.

4.4.1 Dados das Emissões de Poluentes

Este modelo é constituído por duas estrelas distintas, uma relativa aos dados detalhados das emissões de poluentes, outra relativa a dados num maior nível de agregação. O diagrama YAM^2 de nível conceptual mais alto para a estrela das emissões de poluentes detalhadas é apresentado na figura 4.4.

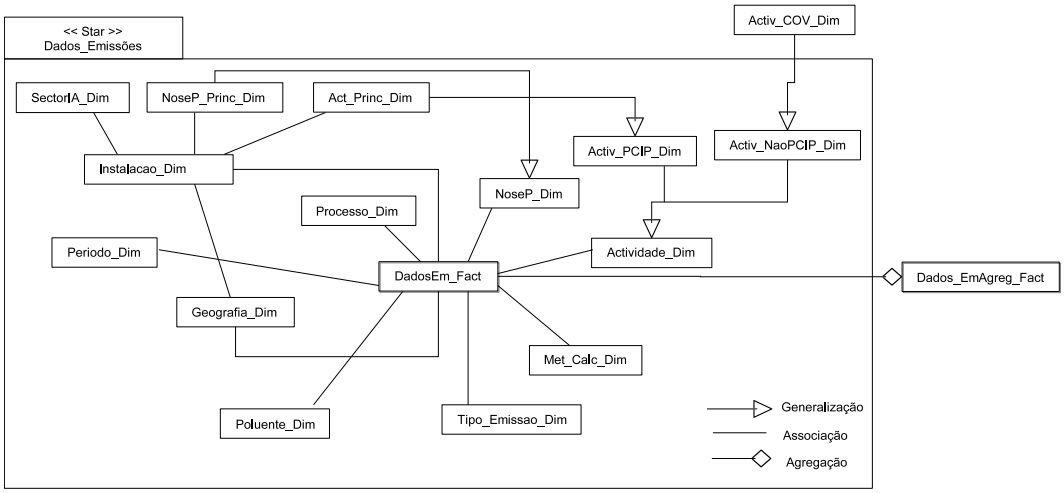

Figura 4.4: Diagrama YAM^2 de nível conceptual mais elevado para a estrela das emissões detalhadas

No detalhe deste diagrama verificamos que existe uma relação de agregação entre dois factos, porque o facto **Dados_EmAgreg** é uma agregação de dados das emissões do facto **DadosEm**. Dado que uma estrela apenas pode conter um facto, o **Dados_EmAgreg** está colocado fora do quadro delimitador da estrela das emissões detalhadas (**Dados-Emissões**), pois pertence a outra estrela (**Dados-EmissõesAgregadas**).

As dimensões que estão relacionadas com os factos de emissões são:

- Instalação_Dim - as emissões referem-se a uma instalação.
- Actividade_Dim e NoseP_Dim - as emissões são um resultado de uma actividade industrial, realizada através de processos industriais. Estão apenas associadas ao facto **DadosEm**, pois o outro facto apresenta os dados de emissões de poluentes agregados ao nível da instalação.
- Poluente_Dim, Tipo_Emissao_Dim e Met-Calc-Dim - os valores de emissões registadas dizem respeito a um poluente, para um determinado tipo de emissão (ar, água-directo, etc.) e são medidas através de um método de cálculo (estimativa, medição, etc.).
- Geografia_Dim - está associada a um único facto (**DadosEm**). No entanto, está associada também à dimensão **Instalação**. Com esta associação da Geografia à Instalação e ao facto **DadosEm** pretende-se caracterizar os dois papéis distintos assumidos pela Geografia: na

dimensão **Instalação** a Geografia representa a própria localização da instalação; no facto das emissões (**DadosEm**) a Geografia indica o local da emissão em si, que se prevê poder ser distinto do local da instalação (embora actualmente as coordenadas precisas dos locais de emissões não sejam recolhidas, apenas são recolhidas as coordenadas da instalação).

- Período_Dim - os valores das emissões referem-se a um determinado período temporal.
- Processo_Dim - a recolha destes dados estão associados a um processo administrativo.

Existem ligações de generalização entre as dimensões **NoseP_Princ** e a dimensão **NoseP**, porque algumas instalações são caracterizadas por um Nose-P principal, e portanto estes serão um subconjunto (no limite, igual ao conjunto total) do universo total de Nose-P existentes.

Da mesma forma, a dimensão **Act_Princ** é a caracterização da instalação por actividade principal. Mas, existe aqui uma diferença relativamente aos Nose-P, pois as actividades principais apenas podem ser actividades PCIP, mas a dimensão **Actividade** tem também actividades não PCIP (**Activ_NaoPCIP**). Por esse motivo foi criada a dimensão intermédia de generalização **Activ_PCIP**. Assim, é possível definir conceptualmente que as actividades principais são um subconjunto das actividades PCIP, e que as actividades PCIP são um subconjunto de todas as actividades (dimensão **Actividade**). A dimensão **Activ_COV** aparece fora da linha delimitadora da estrela porque as actividades COV são um subconjunto das actividades não PCIP (**Activ_NaoPCIP**), estando também incluídas na dimensão **Actividade**. No entanto, não participam nesta estrela, nem neste sub-modelo (só foram apresentadas para se visualizar a generalização).

O **sector-IA** é uma segmentação das instalações que neste momento coincide com as actividades PCIP principais das instalações, prevendo-se que possam vir a ser definidos novos sectores pelo IA, quando houver o conhecimento para fazer esta classificação. Por esse motivo, manteve-se o sector-IA à parte, de forma a permitir uma segmentação independente.

Na figura 4.5 apresenta-se a estrela das emissões agregadas, onde é possível verificar que o facto desta estrela não é indexado pelas dimensões **Actividade**, nem por **NoseP**. Assim, enquanto os dados das emissões detalhadas estão indexados a cada uma das actividades industriais (e a cada um dos processos usados), no facto das emissões agregadas as emissões estão indexadas apenas à instalação. Também devido à agregação efectuada aos dados, este facto não é directamente indexado pela **Geografia**. As restantes dimensões aplicam-se a ambos os factos.

Os dados presentes no facto **Dados_EmAgreg** não podem ser facilmente obtidos a partir dos dados do facto **DadosEm**. Mais do que uma simples agregação, estes dados sofreram também uma ligeira transformação, que é descrita de forma resumida na apresentação do modelo conceptual de nível inferior da estrela **Dados_EmissõesAgregadas**. Caso contrário, poderia

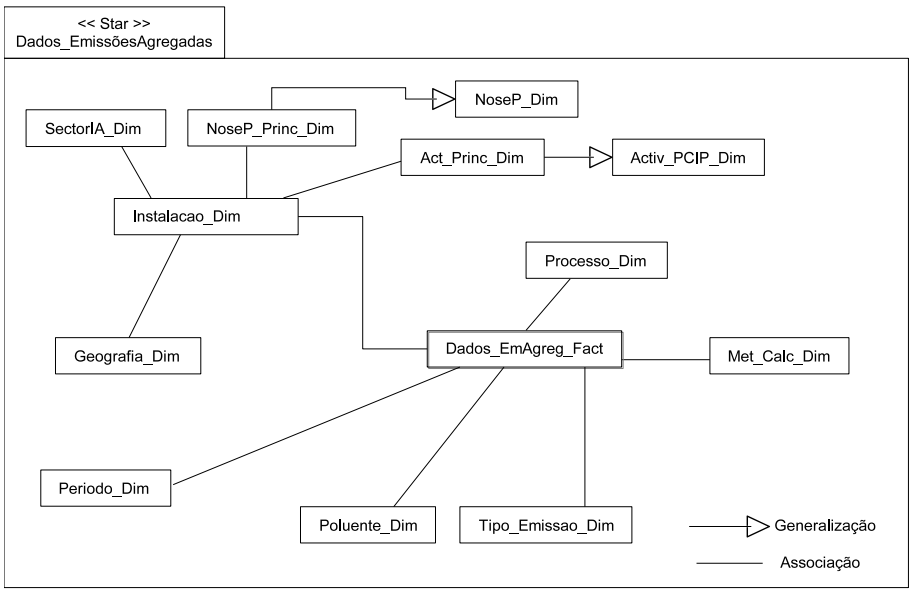

Figura 4.5: Diagrama YAM^2 de nível conceptual mais elevado para a estrela das emissões agregadas

assumir-se que esta estrela seria um dos possíveis agregados a serem previstos para optimizar o desempenho do DW, se tal fosse considerado necessário, e portanto não teria sido incluída no modelo conceptual.

Passando para o nível conceptual intermédio, começamos por apresentar a dimensão **Instalação** na figura 4.6.

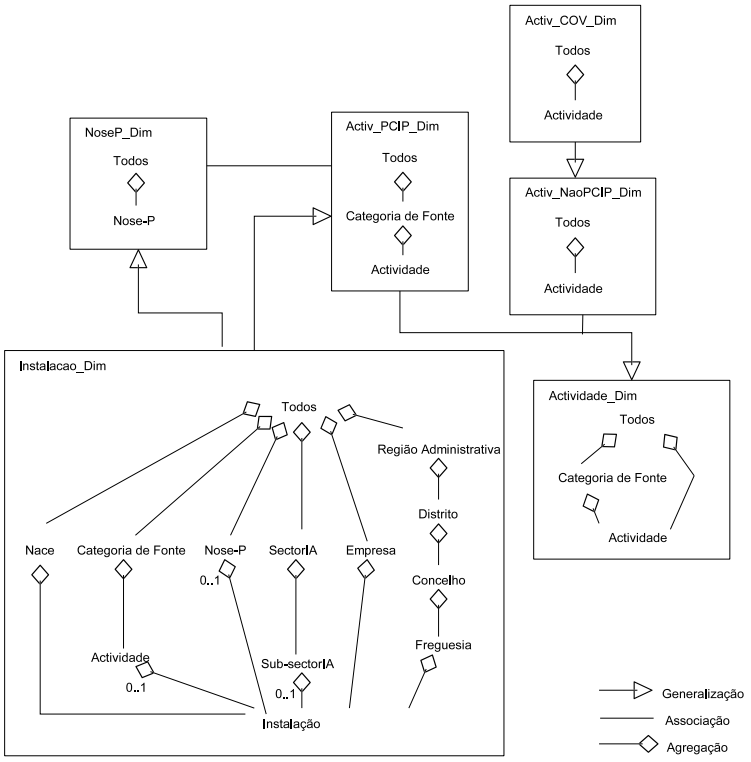

Figura 4.6: Estrutura da dimensão Instalação

Comparando o esquema da dimensão **Instalação** de nível conceptual intermédio com o

que foi representado nos diagramas de nível conceptual mais elevado, verifica-se que foram efectuadas algumas modificações.

A que mais rapidamente se nota é que neste diagrama foram incorporados a actividade principal, o processo Nose-P principal e o sector-IA como hierarquias dentro da dimensão e não à parte. Esta alteração tem a ver com uma tentativa de aproximação do modelo conceptual ao modelo físico e à forma como estes dados são recolhidos na realidade IA, pois são efectivamente considerados atributos classificativos da instalação.

Nos diagramas de nível mais alto achou-se interessante usar a perspectiva de colocar estas hierarquias como dimensões à parte, pois assim é possível representar com maior detalhe a riqueza da informação associada à instalação.

É também de relevar a existência de uma associação entre as dimensões **NoseP** e **Activ_PCIP**, que não tinha sido apresentada antes. Esta associação só é apresentada aqui, e representa a interligação que existe entre os valores destas duas dimensões. Ao nível da UE existem efectivamente tabelas que associam os processos Nose-P às actividades PCIP respectivas, sendo que várias actividades podem estar associadas a diferentes processos e vice-versa. No entanto, estas tabelas não são exaustivas nas associações que indicam, motivo pelo qual esta relação não foi explorada no âmbito deste modelo (podem haver Nose-P utilizados em determinadas actividades PCIP que não estão cobertos pelas tabelas da UE).

Também é possível verificar que existem mais duas hierarquias na dimensão instalação, uma relativa ao código **NACE** e outra à **Empresa**. O código NACE é um agrupamento de Códigos de Actividade Económica (CAE) definido ao nível da UE, onde é utilizado o CAE principal para definir o NACE da instalação. Por exemplo, para o CAE 01131, o código NACE respectivo será 01.13.

A Empresa representa a identificação do proprietário da instalação, sendo que uma empresa pode abranger várias instalações diferentes. A razão pela qual a empresa não foi considerada como dimensão autónoma, mas apenas com um nível da instalação, está relacionada com o facto de o conceito "empresa" não ter ainda grande importância ao nível do IA, embora já tenha uma maior relevância, por exemplo, para a Inspecção Geral do Ambiente e Odenamento do Território(IGAOT) porque, entre outras competências, esta entidade instaura os processos relativos a actos ilícitos na área ambiental, imputáveis à empresa e não à instalação.

As generalizações das actividades PCIP e processos Nose-P já foram explicados no âmbito da estrela de alto nível conceptual dos dados de emissões detalhadas, pelo que não serão novamente descritas. Também aqui a geografia é representada como uma hierarquia de instalação, assumindo totalmente o seu papel de caracterização geográfica das instalações, em vez de ser apresentada como uma dimensão à parte.

As instalações são também classificadas por CAE (embora não se tenha considerado esta

hierarquia, mas apenas a do NACE que se considerou ser mais relevante ao nível da UE), mas para uma dada instalação pode haver vários CAE: um principal e vários secundários. O CAE principal, sendo único, foi incorporado na dimensão **Instalação**. No entanto, esta opção não pode ser utilizada para os secundários, pois cada instalação pode ter um número indefinido de CAE secundários, e que não estão relacionados com as emissões dessa instalação. É portanto numa situação em que uma dimensão tem um número indeterminado de valores.

Uma proposta de implementação desta associação seria a realização de uma *bridge* para os CAE, entre a instalação e o facto, conforme descrito em [KR02]. Esta é uma opção de implementação bastante simples, que consistiria em criar uma tabela associativa entre a dimensão **Instalação** e o facto, conforme se apresenta na figura 4.7.

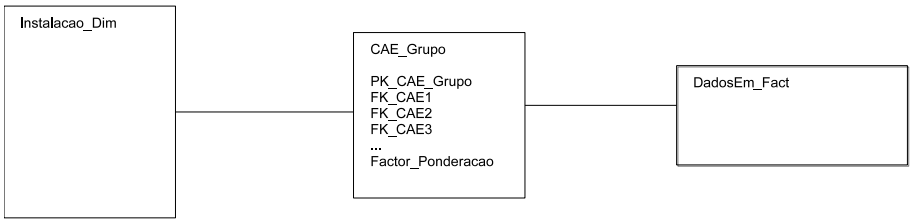

Figura 4.7: Exemplo de uma implementação em *bridge* para as CAE secundárias da instalação

O facto passa a estar associado à chave da tabela de associação, que representa um grupo de CAE. Assim, é possível efectuar qualquer combinação com qualquer número de CAE. Embora na figura não esteja representada, cada chave estrangeira (assinalada com "FK") estará associada à entidade contendo a lista de CAE.

É de salientar que esta implementação não implica qualquer alteração a nenhuma outra entidade, ou seja, é indiferente se a instalação está directamente associada ao facto - opção escolhida no âmbito desta tese - ou se é colocada a tabela associativa para registar o grupo de CAE no meio.

O factor de ponderação que está incluído na tabela do grupo de CAE permite efectuar contas utilizando as CAE: a soma de todos os factores de ponderação das CAE de uma instalação é igual a 1. Desta forma, os valores do facto podem ser ponderados para serem utilizados também ao nível das CAE secundárias, o que de outra forma não seria possível.

Esta opção não foi implementada por várias razões:

1. os CAE secundários das instalações não são muito relevantes para as análises que o IA pretende realizar - o CAE principal deveria ser o que melhor define a actividade da instalação;

2. para esta opção produzir resultados interessantes seria necessário, conforme já referido, definir os factores de ponderação. Estes factores teriam de ter em conta a importância relativa de cada CAE dentro da actividade exercida pela instalação, para o que era

necessário um estudo de detalhe, que exigiria a intervenção e disponibilidade dos técnicos do IA, o que não se verificou ser viável.

Estando terminada a apresentação da dimensão Instalação, passa-se então para a descrição do esquema de nível conceptual intermédio do YAM^2 relativo aos dados das emissões detalhadas, apresentado na figura 4.8.

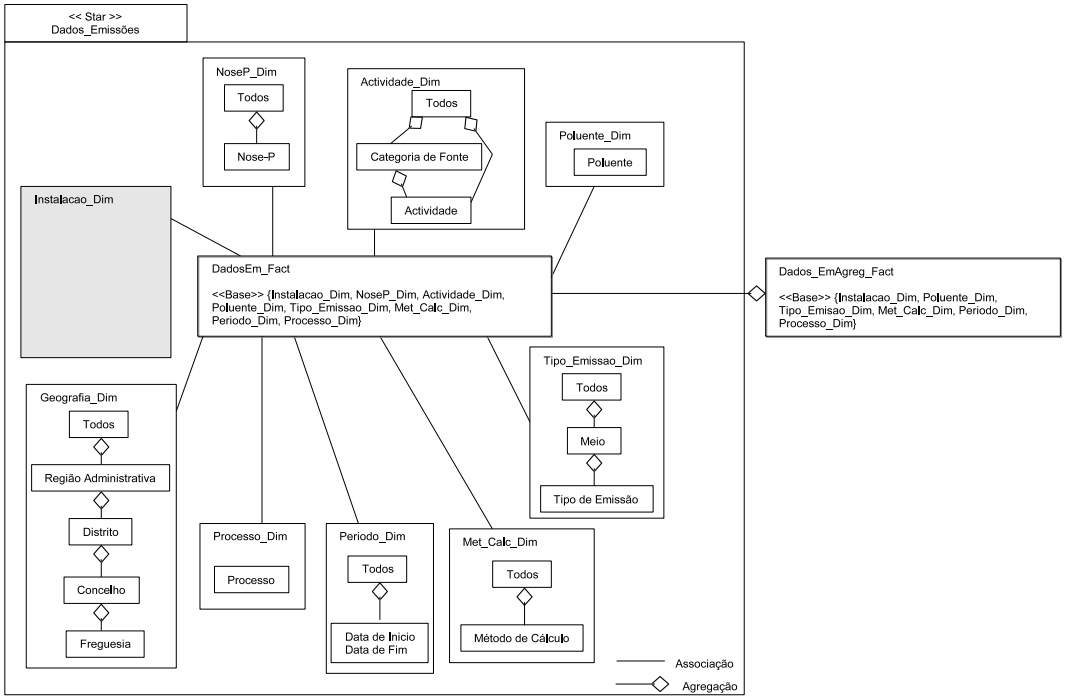

Figura 4.8: Diagrama YAM^2 de nível conceptual intermédio para os dados das emissões detalhadas

Na figura 4.9 apresenta-se o esquema de nível conceptual intermédio do YAM^2 relativo aos dados das emissões agregadas. Dada a similaridade entre estes dois esquemas, eles serão explicados em simultâneo.

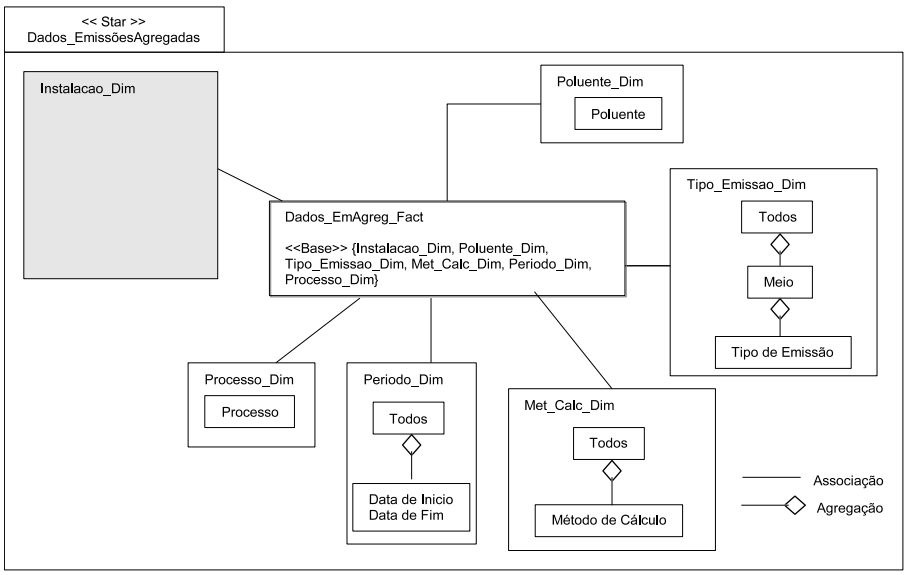

Figura 4.9: Diagrama YAM^2 de nível conceptual intermédio para os dados das emissões agregadas

Conforme já descrito, a grande alteração de nomenclatura relativamente à descrita no YAM[2], nestes diagramas, é o facto de a dimensão **Instalação** ser apresentada como uma caixa sombreada, que já foi descrita em detalhe e portanto não se considerou relevante descrevê-la novamente, simplificando bastante o modelo conceptual. Pelo mesmo motivo as associações da dimensão **Instalação** para as dimensões **NoseP** e **Actividade** não são novamente apresentadas. Assim, a dimensão **Geografia** é apenas apresentada no seu papel caracterizador das emissões, associada ao facto **DadosEm** (já não é apresentada associada à instalação).

As dimensões **Met_Calc**, **Tipo_Emissao** e **Poluente** caracterizam cada emissão. Chama-se a atenção para o caso particular da dimensão **Poluente**, que não tem a agregação da hierarquia para o nível "Todos´´. Não faz sentido análises de emissões somando valores de vários poluentes, devido às distintas ordens de grandeza dos valores dos vários poluentes. Embora todos os poluentes sejam medidos em quilogramas, pode dar-se o caso de um ter valores na ordem das milésimas, e outros valores na ordem das centenas de milhão (por exemplo, o CO_2), portanto a soma dos poluentes gera resultados pouco interessantes, ou mesmo inconsistentes. No caso específico de ser utilizada a medida "Quantidade emitida em limiares" (igual à divisão da quantidade emitida do poluente pelo limiar do poluente), poderão ser efectuadas somas de quantidades de vários poluentes, pois o resultado desta medida é a quantidade emitida do poluente expressa em número de limiares, o que terá o efeito de uma normalização de valores. Mas como esta é uma situação muito particular, considera-se para todos os efeitos que a agregação para todos os poluentes não deverá ser representada como possível.

Também é possível verificar a existência de duas hierarquias distintas para a dimensão **Actividade** e que estão relacionadas com a existência de valores de emissões para actividades não-PCIP. Para as actividades não-PCIP, não será possível a agregação pelas categorias de fonte, pois estas são agrupamentos de actividades PCIP.

Relativamente às bases dos factos [1], estas são construídas tendo em conta o conjunto das dimensões necessárias e suficientes para identificar univocamente cada valor do facto. Assim, verifica-se que no facto **DadosEm** a dimensão **Geografia** não faz parte da base, porque neste momento a geografia das emissões não é recolhida, embora tivesse sido prevista no modelo para maior completude (acresce que nunca foram relatadas duas emissões distintas devido à distribuição geográfica dos locais de emissão de poluentes). Todas as outras dimensões estão incluídas na base do respectivo facto. A dimensão **Período** poderá ser uma dimensão de agregação sempre que estejamos perante períodos disjuntos, ou seja, não exista sobreposição de dados relativamente a um dado período, ou que sejam identificadas as versões mais recentes dos dados para não existir sobreposição.

[1] Conforme descrito na secção 2.6, a base é o conjunto mínimo de níveis necessários e suficientes para identificar univocamente uma célula

Verifica-se novamente a existência da associação de agregação entre os dois factos, indicando que o facto **Dados_EmAgreg** é obtido a partir da agregação de valores presentes no facto **DadosEm**. Mais concretamente, os valores do facto **Dados_EmAgreg** são obtidos agregando os dados do facto **DadosEm** para as **Actividades**, **NoseP** e **Geografia**, que portanto são dimensões não relacionadas com o facto agregado.

Passa-se então à apresentação do modelo conceptual de nível inferior para os dados das emissões. É de ressaltar que não se pretende nesta secção apresentar detalhadamente o modelo físico do modelo, mas apenas o conceptual. Portanto, não se considera relevante nem essencial a apresentação total de todos os descritores das dimensões. os descritores considerados necessários para a compreensão das medidas dos factos são apresentados na tabela 4.1, pertencendo todos eles à dimensão **Período** (foram consideradas várias formas de medição do período, que são depois utilizadas nas medidas criadas).

Tabela 4.1: Descritores necessários para a compreensão das medidas dos factos

Dimensão onde pertence	Descritor	Significado
Período	Duração em semestres	Quantos semestres são abrangidos pela informação prestada
Período	Duração em trimestres	Quantos trimestres são abrangidos pela informação prestada
Período	Duração em meses	Quantos meses foram abrangidos pela informação prestada
Período	Duração em dias	Quantos dias são abrangidos pela informação prestada
Período	Tipo de duração	Um dos seguintes valores: anual, semestral, trimestral, mensal

Na tabela 4.2 apresentam-se as medidas existentes na estrela **Dados_Emissoes**.

A informação apresentada nesta tabela para cada medida é a seguinte:

- Nome da medida.
- Fórmula de cálculo, que se não estiver preenchida é porque se trata de um valor recolhido, e não calculado.
- Dimensões de agregação, ou seja, o conjunto de dimensões sobre as quais a medida é agregável (se estiver preenchida com um '-', trata-se de um valor não agregável). Por exemplo, sendo uma medida $\underline{m}$ indexada por uma dimensão $\underline{d}$, diz-se que $\underline{m}$ é agregável, através de um operador de agregação $\underline{G}$, sobre a dimensão $\underline{d}$, se for possível agregar vários valores de $\underline{m}$ associados a distintos elementos de $\underline{d}$ num único valor $\underline{G}(\underline{m})$.

Tabela 4.2: Medidas do facto **Dados_Em**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Quantidade_emitida	-	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão, Actividade, NoseP	Sum, Average, Count, Max, Min
Limiar_poluente	-	-	-
Média_semestral	quantidade/(duração em semestres)	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão, Actividade, NoseP	Max, Min
Média_trimestral	quantidade/(duração em trimestres)	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão, Actividade, NoseP	Max, Min
Média_mensal	quantidade/(duração em meses)	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão, Actividade, NoseP	Max, Min

- Funções de agregação suportadas pela medida (preenchidas com um '-' no caso de valores não agregáveis).

Analisando a tabela, verifica-se que as dimensões **Poluente** e **Processo**, que estão associadas à tabela de factos, não estão indicadas como dimensões de agregação. O motivo pela qual os dados não podem ser agregados pela dimensão **Poluente** já foi explicado no âmbito da descrição do diagrama de nível conceptual intermédio. Quanto à dimensão **Processo**, assume-se que não será possível agregar dados de diferentes processos administrativos, porque esta dimensão foi criada apenas com o objectivo de identificar univocamente os dados, para efeitos de rastreabilidade. Além disso, existem duas possibilidades diferentes para a recolha de dados, cada uma delas influenciando a questão da agregação ao longo de processo.

1. Pode-se permitir que sejam recolhidos vários conjuntos de dados, associados a processos administrativos diferentes, mas para o mesmo período (já aconteceu durante uma recolha de dados em 2002, onde era permitida periodicidade trimestral na prestação de dados, e depois uma última informação anual abrangendo os 4 trimestres), caso em que a agregação ao longo do processo provocaria duplicação de valores. Neste caso, é necessário identificar quais os processos que devem ser contabilizados na própria dimensão **Processo**, para permitir a utilização correcta de dados que existam em duplicado. Se forem identificadas todas as versões mais actuais dos processos, a dimensão Período poderá ser uma dimensão de agregação, pois não existirá sobreposição de dados relativos a um mesmo período.

2. Ou, pode-se assumir que não há redundância de dados para um mesmo período e, neste caso, agregar ao longo dos processos teria o mesmo efeito da agregação ao longo do período, portanto não produzia um resultado interessante.

Considera-se que as funções de agregação da medida "Quantidade emitida´´ são a soma, a média e a contagem de valores, além dos operadores máximo e mínimo. As medidas de quantidades de emissão são as únicas que têm como função de agregação prevista a contagem, porque o número de emissões reportadas tem algum significado, principalmente quando comparado com o número de instalações que fizeram essas emissões. Considera-se que também faz sentido obter uma média de emissões, por exemplo para calcular médias de emissões por instalação.

A medida "Limiar_Poluente´´, tal como o nome indica, refere-se aos limites dos poluentes definidos pela UE, utilizados para filtrar as emissões que terão de ser reportadas à UE (acima do limiar), daquelas que não serão reportadas (inferiores ao limiar). Dado que os limiares são quantidades fixas, não faz sentido agregá-los. Os limiares dos poluentes foram introduzidos nesta tabela para permitir calcular novas medidas, como por exemplo as quantidades emitidas expressas em limiares.

As medidas que já são apresentadas sob a forma de médias não se consideram agregáveis, excepto através da utilização dos operadores máximo e mínimo, pois as médias estão directamente relacionadas com o número de valores que estão a contabilizar, que não é controlável. Portanto, não é possível garantir que a média de médias, num nível mais agregado, seja equivalente ao cálculo da média utilizando os dados das quantidades emitidas para esse nível de agregação. Estas medidas utilizam os descritores da dimensão **Período** (duração em semestres, duração em trimestres, duração em meses), anteriormente apresentados como parte da sua fórmula de cálculo.

O facto **Dados_EmAgreg** da estrela "Dados_EmissoesAgregadas" tem as medidas apresentadas na tabela 4.3.

Nesta tabela está prevista a existência simultânea de dois tipos de dados:

- Os dados que foram reportados pelos operadores, que aqui já foram previamente agregados - medida "Quantidade_reportável´´. Na fórmula de cálculo está dada a indicação desta medida ser obtida a partir da soma da "Quantidade_emitida´´ do facto **Dados_Em**, e à frente (entre {}) são apresentadas as dimensões pelas quais esta medida está agregada: **Actividade**, **NoseP** e **Geografia**.

- Os dados que são convertidos pelos técnicos do IA, e que serão reportados à UE - medida "Quantidade_reportada´´. A fórmula de cálculo é apresentada como "ConversãoUE´´, dado que se pretende generalizar as operações de conversão efectuadas sobre os dados. A conversão tem 4 regras, relacionadas com a forma como os dados são relatados à UE: no âmbito do EPER só se pretende que sejam relatados os dados relativos aos maiores poluidores de cada país. Portanto, as primeiras 3 regras da conversão fazem os cálculos relativamente aos limiares dos poluentes, existindo um tratamento diferencial das emis-

Tabela 4.3: Medidas do facto **Dados_EmAgreg**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Quantidade_ reportada	ConversaoUE(Quantidade_ reportavel)	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão	Sum, Average, Count, Max, Min
Quantidade_ reportavel	Sum(Dados_Em_ Fact:Quantidade_ Emitida) {Actividade, NoseP, Geografia}	Instalacao, Geografia, Período, Metodo_Calculo, Tipo_emissão	Sum, Average, Count, Max, Min
Limiar_ poluente	-	-	-
Média_ semestral	quantidade/ (duração em semestres)	Instalacao, Geografia, Metodo_Calculo, Tipo_emissão	Max, Min
Média_ trimestral	quantidade/ (duração em trimestres)	Instalacao, Geografia, Metodo_Calculo, Tipo_emissão	Max, Min
Média_ mensal	quantidade/ (duração em meses)	Instalacao, Geografia, Metodo_Calculo, Tipo_emissão	Max, Min

sões associada a actividades presentes na lista PCIP, das outras:

1. Para as emissões associadas a actividades PCIP - soma-se as emissões da instalação ao nível do "Meio" (somando em separado "água-directo" com "água-indirecto", e considerando o "ar" sozinho). Se o valor de cada "Meio" para cada instalação ultrapassar o limiar, estes dados serão relatados à UE (e aparecerão na medida "Quantidade_reportada" do facto **Dados-EmAgreg**).
2. Para as emissões associadas a actividades não PCIP - Se o Nose-P associado à emissão corresponder a um processo de Estação de Tratamento de Águas Residuais (ETAR), adiciona-se as emissões à actividade principal da instalação e faz-se as contas como se fossem também emissões associadas a actividades PCIP, e volta a executar a regra 1.
3. Para as emissões associadas a actividades não PCIP - Se o Nose-P não corresponder a uma ETAR, verifica-se se a soma das emissões das actividades não PCIP da instalação ultrapassa os 10% das emissões totais da instalação, para aquele "Meio". Em caso afirmativo, estas emissões são também tratadas como se estivessem associadas a actividades PCIP e volta a executar a regra 1. Caso contrário, as emissões são ignoradas.
4. Para qualquer emissão que tenha ultrapassado o limiar - os valores são formatados a 3 dígitos significativos (por exemplo 123456 passa para 123000).

As medidas relativas às médias são semelhantes às já descritas na tabela relativa ao facto **Dados_Em**, pelo que não são novamente apresentadas. As dimensões de agregação não in-

cluem novamente a dimensão **Poluente**, pelos motivos já referidos anteriormente, mas também não incluem a **Actividade** e o **NoseP**, dado que este facto não é indexado por essas duas dimensões.

4.4.2 Dados de Produção

Apresenta-se na figura 4.10 o modelo conceptual de alto nível para este sub-modelo, sendo possível verificar que nesta estrela não existem as dimensões **Met_Cálc**, **Tipo_emissao** e **Poluente** (dimensões relacionadas com emissões), mas continuam a existir as dimensões **Actividade** e **NoseP**, que também indexam os dados de produção.

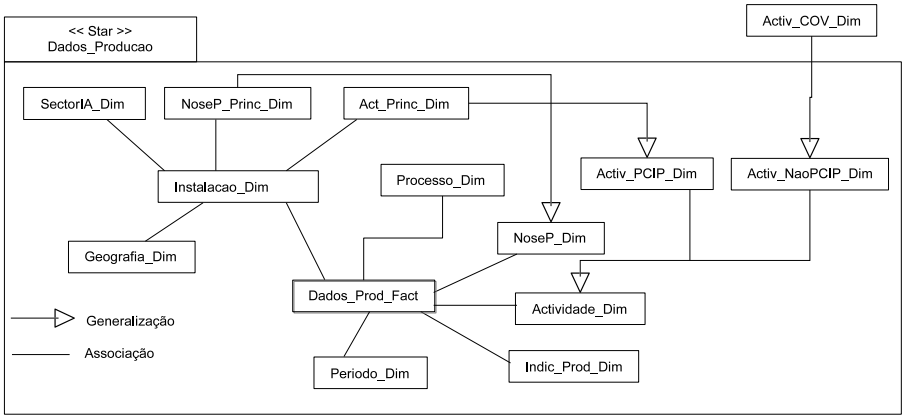

Figura 4.10: Diagrama YAM^2 de nível conceptual superior para os dados de produção

Isto porque os dados de produção estão relacionados com o produtos desenvolvidos pelas instalações, e que foram criados no âmbito de uma actividade industrial, através de um processo Nose-P. Estes dados consideram apenas o resultado efectivo da produção da instalação, e portanto não estão aqui a ser contabilizadas as produções residuais como as emissões de poluentes, motivo pelo qual nesta estrela não aparecem as dimensões associadas a emissões.

Verifica-se que aqui a geografia existe apenas como caracterizadora da instalação, porque se considera que a produção tem de estar na instalação e não aparecer numa localização independente.

Mais uma vez, as actividades COV encontram-se fora da estrela, porque a produção de uma instalação não está relacionada com estas actividades.

Para o modelo conceptual de nível intermédio, temos o diagrama apresentado na figura 4.11.

Aqui salienta-se a dimensão **Indic_Prod**, que corporiza os vários indicadores de produção incluídos neste facto e as suas unidades, não tendo a agregação para o nível "Todos´´. Na realidade, sendo os indicadores de produção valores tão variados como:

- Capacidade Instalada - capacidade máxima da instalação para a realização da sua actividade, no período a que dizem respeito os dados.

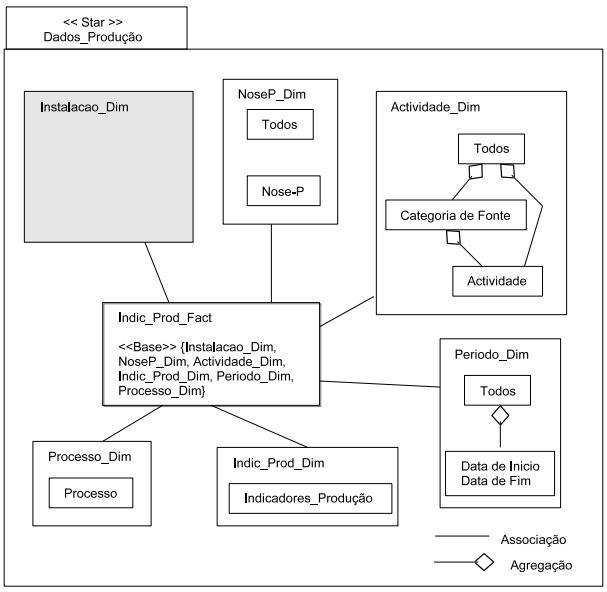

Figura 4.11: Diagrama YAM^2 de nível conceptual intermédio para os dados de produção

- Capacidade efectivada - é a capacidade efectivamente utilizada na instalação no período a que dizem respeito os dados.
- Volume de produção - Quantifica a produção da instalação no período a que dizem respeito os dados.

E tendo em conta que, para cada um destes valores, podem ser recolhidos dados em unidades diferentes em simultâneo (caso em que se criará um indicador diferente para cada unidade), a agregação de dados provenientes de vários indicadores provocaria valores inconsistentes.

Mais uma vez, a dimensão **Actividade** apresenta as duas hierarquias alternativas correspondentes às actividades PCIP e não PCIP, pois os dados de produção também podem ser recolhidos para estes dois tipos de actividades. Quanto ao nível conceptual inferior, a descrição do facto **Dados_Prod** é apresentada na tabela 4.4.

Tabela 4.4: Descrição do facto **Dados_Prod**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Valor_ efectivo	ConversaoIA (Valor_Op)	Instalacao, Período, Actividade, NoseP	Average, Max, Min
Valor_Op	-	Instalacao, Periodo, Actividade, NoseP	Average, Max, Min
Unidade_Op	-	-	-

A medida ”Valor_op´´ representa o valor efectivamente reportado pelo operador da insta-

lação, ou seja, antes de qualquer verificação ou conversão efectuada pelos técnicos do IA.

A ”Unidade_Op´´, embora tenha sido incorporada no facto, não é uma medida, pois este valor identifica as unidades utilizadas pelo operador na sua informação ao IA. No entanto, dado que se assume que estes valores reportados pelo operador podem não estar normalizados (quer em termos de unidades, quer em termos de correcção de valor), eles são adicionados ao facto para completude da informação, mas a medida ”Valor_efectivo´´ é que contém os valores que realmente são relevantes para as análises. A fórmula de cálculo ”ConversaoIA´´ identifica que esta medida contém o valor verificado e normalizado pelos técnicos do IA, e portanto é com este valor que se poderão obter resultados mais correctos e fiáveis.

A soma não faz parte das funções de agregação porque, dependendo de qual o indicador de produção considerado, poderá não fazer sentido somar os seus valores. Por exemplo, para uma dada instalação não faz sentido somar as suas capacidades efectivadas ou instaladas ao longo do tempo, mas fará sentido somar o seu volume de produção. Caso a caso é que terá de ser avaliado se o indicador pode ou não ser agregado através de soma.

4.4.3 Dados de Funcionamento

Apresenta-se na figura 4.12 o diagrama de nível conceptual superior para os dados de funcionamento.

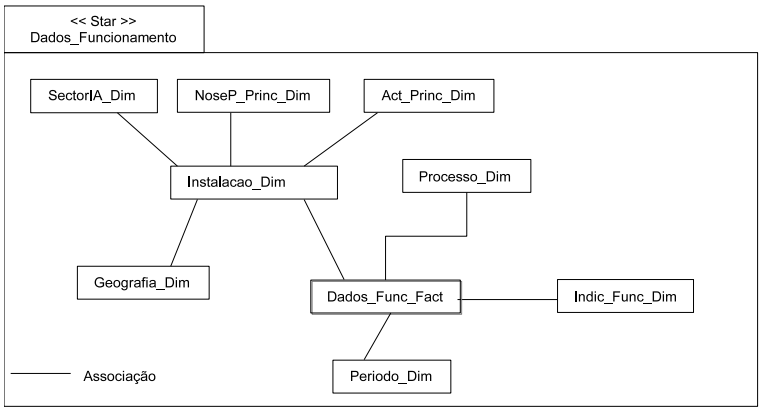

Figura 4.12: Diagrama YAM^2 de nível conceptual superior para os dados de funcionamento

As principais diferenças deste esquema relativamente ao anterior residem nas dimensões **Actividade** e **NoseP**, que não estão presentes neste esquema. Os dados de funcionamento estão directamente relacionados com as instalações, e não com as actividades das mesmas ou os Nose-P utilizados, embora, conforme já foi referido na apresentação dos conceitos, fosse interessante se estivessem.

Neste sub-modelo, à semelhança da estrela “Dados_Produção” também foi definida uma dimensão que identifica os vários indicadores de funcionamento, denominada **Indic_Func**. Alguns exemplos, destes indicadores são:

- Número médio de empregados
- Número de horas de funcionamento no período
- Número de turnos de laboração
- Duração dos turnos de laboração
- Número de semanas de laboração por ano
- Número de dias de laboração por semana
- Número de horas de laboração por dia
- Número de trabalhadores
- Consumo médio anual de água da rede pública
- Consumo total anual de água da rede pública
- Consumo total anual de água
- Consumo médio anual de energia eléctrica
- Potência instalada, em kVa e kW - o kVA é uma medida de potência eléctrica, chamada potência aparente, os kW representam a potência activa, ou seja, a energia consumida.
- Valor da intensidade energética - quantificação da intensidade energética da instalação (em energia consumida por unidade de produto acabado).

Também nesta dimensão, dada a diversidade dos vários indicadores envolvidos, se assume que não pode haver agregação por todos os indicadores, como se pode ver pelo diagrama de nível intermédio apresentado na figura 4.13. Mais uma vez, as unidades associadas a cada indicador encontram-se referidas na dimensão.

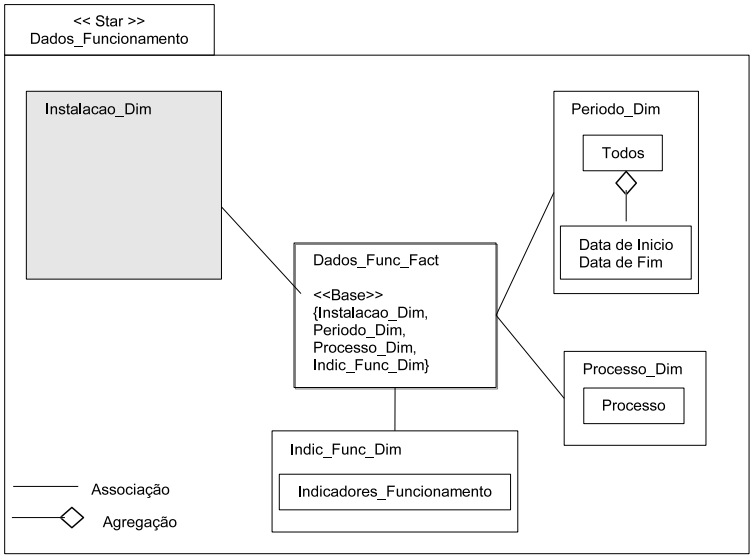

Figura 4.13: Diagrama YAM^2 de nível conceptual intermédio para os dados de funcionamento

O diagrama de baixo nível conceptual do facto **Dados_Func** é apresentado na tabela 4.5.

Tabela 4.5: Medidas do facto **Dados_Func**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Valor_ efectivo	ConversaoIA (Valor_OP)	Instalacao, Periodo	Average Max, min
Valor_Op	-	Instalacao, Periodo	Average, Max, min
Unidade_Op	-	-	-

A base de agregação desta tabela de factos é constituída por todas as dimensões que a indexam, excepto a dimensão **Indic_Func**, que, tal como já foi referido anteriormente para os dados de produção, não tem o nível de agregação "Todos´´ porque não faz sentido juntar dados pertencentes a vários indicadores de funcionamento. A medida "Valor_op´´ representa novamente o valor efectivamente reportado pelo operador da instalação, e o atributo "Unidade_Op" identifica as unidades utilizadas pelo operador na sua informação. Mais uma vez, a medida "Valor_efectivo" é que contém os dados mais relevantes para a realização de análises.

Também estes indicadores não têm a função de agregação "soma", pois considera-se que para alguns indicadores poderá não fazer sentido somar os seus valores (a lista de indicadores apresentada não é exaustiva).

4.4.4 Dados de COV

São chamados Compostos Orgânicos Voláteis (COV) todos os compostos orgânicos resultantes da actividade humana, à excepção do metano[2], que possam produzir oxidantes fotoquímicos por reacção com óxidos de azoto, na presença de luz solar [IAP] e incluem substâncias como a acetona, o metanol, o etanol, o tolueno, o diclorometano e outros hidrocarbonetos clorados, ou misturas destes compostos, assim como compostos aromáticos. [LS]

Uma grande parte da emissões de COV para a atmosfera deve-se à utilização de solventes em tintas, vernizes e produtos de retoque de veículos. As principais abordagens para a diminuição das emissões de COV incluem: substituição dos produtos de base solvente por produtos aquosos; redução do teor de solventes nos produtos de base; redução do teor de solventes nos produtos aquosos. Por esse motivo, um dos principais dados que se solicitam aos operadores sobre os COV é a quantidade de solventes que consomem. Este sub-modelo é apresentado na figura 4.14.

As dimensões aplicáveis a esta estrela são a **Instalacao**, **Periodo**, **Processo** e **Actividade**. As actividades aparecem como uma generalização das actividades Não-PCIP (**Activ_NãoPCIP**),

[2]O metano resulta essencialmente de fontes naturais ou semi-naturais como a fermentação, extracção e transporte de gás natural, produção animal. O metano não é tóxico e é praticamente inerte do ponto de vista fotoquímico, e foi até há bem pouco tempo incluído nos COV. No entanto, considerando as suas concentrações não negligenciáveis na atmosfera e a sua contribuição para o efeito de estufa, o metano tem sido considerado separadamente, falando-se actualmente em Compostos Orgânicos Voláteis Não Metânicos. [dAeOdT04]

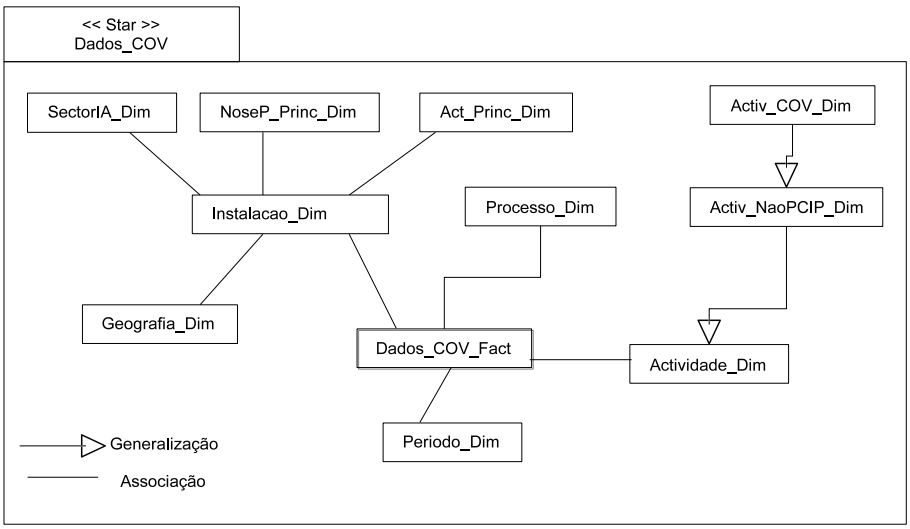

Figura 4.14: Diagrama YAM^2 de nível conceptual superior para os dados de COV

que por sua vez generalizam as actividades COV (**Activ_COV**), actividades associadas à emissão de COV (por exemplo, "B) Actividades de revestimento a) Veículos Cabines de camiões abrangidos pelas categorias N2,N3"). Estas já estão incluídas dentro da linha delimitadora da estrela, uma vez que são efectivamente utilizadas neste sub-modelo.

Os consumos solicitados não estão indexados ao tipo de solvente consumido, até porque podem dizer respeito a vários solventes em conjunto. Por este motivo, não foi criada uma dimensão "solvente" para caracterizar o solvente consumido. Assim, estes dados indicam apenas as quantidades consumidas, para cada uma das actividades COV da instalação.

A dimensão **Geografia** também não indexa este facto, pois assume-se que os solventes deverão ser utilizados dentro da instalação, sendo emitidos no local onde são utilizados (não existirá transporte destas emissões). Assim, a localização da instalação é suficiente para indicar a localização das emissões de COV.

No diagrama de nível conceptual intermédio, apresentado na figura 4.15, as actividades COV já se encontram incluídas dentro da dimensão **Actividade**, para melhor representar o modelo real, já que a dimensão **Actividade** inclui também estas actividades. A separação destas duas dimensões que foi feita no diagrama de nível conceptual superior teve como objectivo enriquecer a informação transmitida pelo modelo conceptual.

Para identificar univocamente cada registo de COV (consumo de solventes) basta considerar as dimensões **Instalação**, **Actividade**, **Período** e **Processo**.

Ao nível das dimensões de agregação da tabela de factos **Dados_COV**, apresentadas no diagrama de nível conceptual inferior (tabela 4.6), não aparece a dimensão **Processo**, pelo mesmo motivo já explicado nos anteriores diagramas.

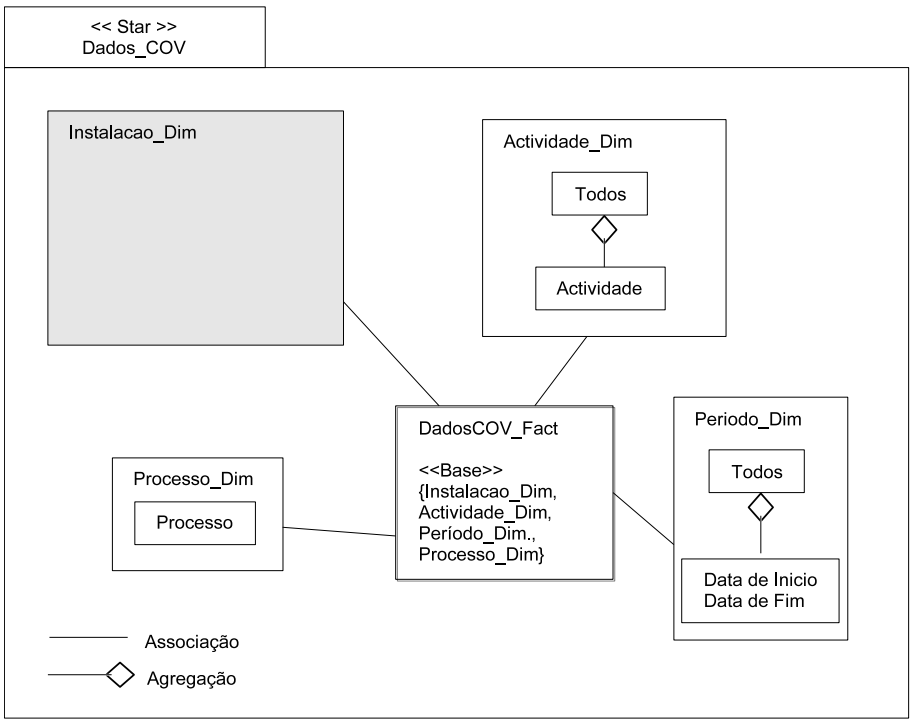

Figura 4.15: Diagrama YAM^2 de nível conceptual intermédio para os dados de COV

Tabela 4.6: Medidas do facto **Dados_COV**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Consumo	-	Instalacao, Geografia, Período, Actividade	Sum, Average, Count, Max, Min

4.4.5 Dados de descarga de águas residuais

O formulário PCIP abrange as descargas de águas residuais que são efectuadas pelas instalações. Este formulário não detalha qual o conteúdo (em termos de substâncias) das descargas, mas apenas as quantidades descarregadas e suas periodicidades, e o tipo de encaminhamento dado ao resultado da descarga.

O diagrama de nível conceptual superior desta estrela é apresentado na figura 4.16.

Este diagrama abrange vários tipos de descarga de águas residuais, que são diferenciadas conforme a sua origem, tipo de regime de descarga e destino. Quanto à origem, as descargas de águas residuais podem ser classificadas como "Domésticas", "Pluviais", "Industrial" e "Doméstico + Industrial". Os tipos de regime de descarga podem ser "contínuas", "descontínuas", "esporádicas" e "potenciais" (quando se consideram os derrames acidentais, esvaziamento de reservatórios, etc.).

A origem e o tipo das descargas foram incluídas na dimensão única **Caract_Descarga**, que caracteriza a descarga quanto a estes dados. Estas duas caracterizações não foram separadas devido à diminuta lista de valores que cada uma contém, e porque apresentam uma forte cor-

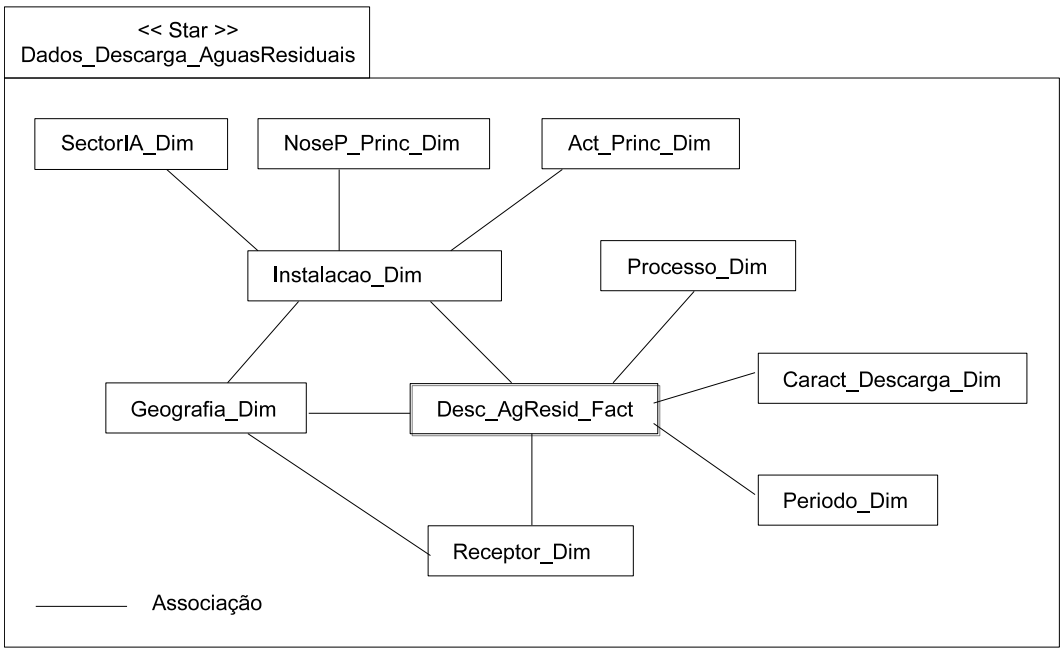

Figura 4.16: Diagrama YAM^2 de nível conceptual superior para os dados de descargas de águas residuais

relação entre elas.

Quanto ao destino, este é incluído numa dimensão própria pois existem mais dados caracterizadores do próprio receptor além do seu tipo, nomeadamente o nome, a bacia hidrográfica a que pertence e alguns dados adicionais, distintos conforme o tipo de receptor. Foi opinião dos técnicos do IA que esta seria uma dimensão de análise relevante, motivo pelo qual se optou por incluir todos os receptores na mesma dimensão, e não separar os tipos de receptores. Este é diferenciado através do tipo de receptor a que se destina a descarga e que tem os seguintes valores: "águas de superfície", "solo / águas subterrâneas" e "sistemas de drenagem colectivos" (ETAR).

No tipo de receptores "águas de superfície" são solicitados dados relativos ao caudal: caudal médio anual, caudal de ponta e de estiagem. Embora sendo constituídos por valores numéricos, considerou-se que os dados relativos ao caudal do receptor não deveriam ser agregados (pelo menos no âmbito dos dados abrangidos por este sub-modelo), mas serviriam para caracterizar o receptor. Por este motivo, não foi criada uma tabela de factos para estes dados, tendo sido incluídos na dimensão.

No tipo de receptor "solo/águas subterrâneas", é solicitado o tipo de solo (Argiloso, Arenoso, etc.), o uso do solo receptor (Cultura Hortícola, Cultura Agrícola Não Hortícola, Floresta, Solo Não Cultivado, etc.), a área (em hectares), e o titular do terreno.

No tipo de receptor "sistemas de drenagem colectivos" é solicitado qual o tipo de sistema (ETAR Municipal, ETAR Industrial, ETAR Mista, Ausência de ETAR de Destino), o nome do sistema, a entidade detentora do sistema e a entidade transportadora.

É possível verificar que a dimensão **Geografia** está a indexar directamente o facto **Desc_AgResid**, além da indexação à dimensão **Instalacao**). Isto porque, como acontece para as emissões de poluentes, se assume que os locais da descarga podem estar localizados fora da instalação (através de condutas de transporte), e portanto é necessário identificar a área afectada pelas descargas. Poderia ter sido criada uma dimensão específica para identificação dos pontos de descarga mas, não tendo sido possível identificar mais dados para a sua caracterização além das suas coordenadas, esta seria uma dimensão autónoma, mas sem um conjunto de dados caracterizadores.

É necessário considerar também que um local de descarga poderá ser usado por mais do que uma instalação, podendo cada instalação usar vários pontos de descarga, portanto não existe a possibilidade de afectar cada ponto de descarga à sua instalação (existe uma relação de vários para vários entre estas duas dimensões), portanto a dimensão "pontos de descarga" seria apenas associada à tabela de factos. Tendo em conta estas considerações, optou-se por incorporar as coordenadas directamente na tabela de factos (permitindo a posterior associação à **Geografia**), e prescindir da existência da dimensão específica "ponto de descarga". Mais uma vez, a geografia dos pontos de descarga não é actualmente recolhida, mas foi prevista no modelo por uma questão de maior abrangência.

A geografia indexa também o receptor, pois a sua localização em termos administrativos pode ser relevante. É necessário ter em conta que esta associação não se aplica a todos os tipos de receptores (por exemplo, não se aplica aos receptores do tipo "águas de superfície"), conforme será explicado a seguir.

Passando ao esquema de nível conceptual intermédio apresentado na figura 4.17, pode-se verificar que se prevê a inclusão da bacia hidrográfica na dimensão **Receptor**, conforme já referido.

Por definição, a bacia hidrográfica é o conjunto de meios hídricos - aquáticos - cujos cursos ou leitos se interligam, e é um conjunto de terras banhadas por um rio principal e seus tributários - afluentes, subafluentes - pelo que uma ETAR não se encontra incluída na classificação de bacia hidrográfica. A inclusão deste nível hierárquico tem como objectivo permitir a análise dos dados relativos a descargas ao longo das várias bacias hidrográficas existentes, um eixo geograficamente relevante.

A razão de não se incluir a bacia hidrográfica directamente na geografia está relacionada com a aplicabilidade: a bacia hidrográfica não é relevante para os outros dados analisados, apenas nas descargas de águas residuais, motivo pelo qual é incluída apenas nesta dimensão. Por vezes o agrupamento por este nível pode não ser aplicável, como por exemplo quando os receptores são ETAR, sendo por isso apresentada a hierarquia alternativa na dimensão **Receptor**. Para as ETAR será porventura mais relevante usar a associação à dimensão **Geografia** para

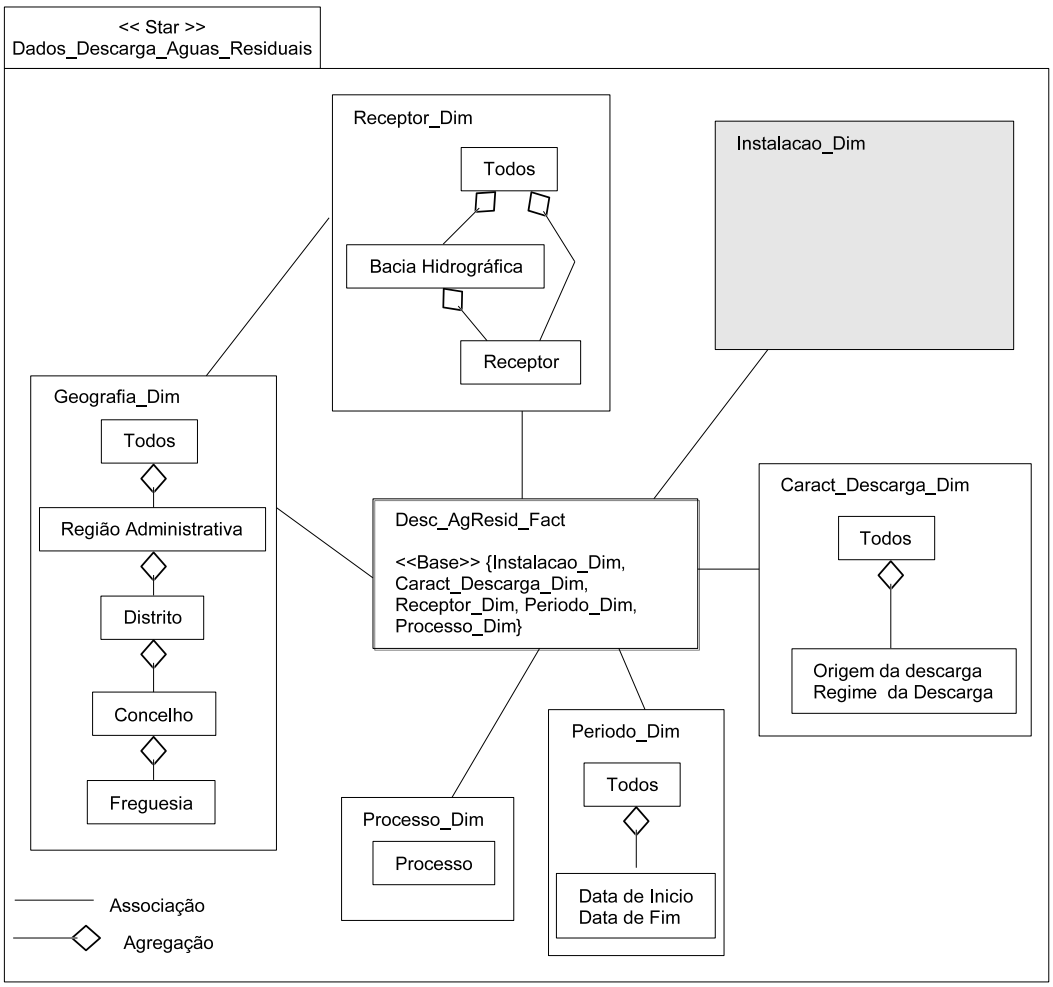

Figura 4.17: Diagrama YAM^2 de nível conceptual intermédio para os dados de descargas de águas residuais

indicar a sua localização (administrativa), e a bacia hidrográfica será utilizada para os outros tipos de receptores, aos quais a dimensão geografia não é aplicável (quando se considera cursos de água, estes podem atravessar várias zonas geográficas administrativas - concelhos, distritos, etc.).

Quanto à base de identificação unívoca dos valores do facto, verificamos que apenas a **Geografia** não está incluída na base, pois a identificação da instalação, a caracterização da descarga e o receptor serão à partida suficientes para identificar a descarga.

Relativamente ao esquema de nível conceptual inferior deste esquema, o seu conteúdo é o apresentado na tabela 4.7.

As primeiras três medidas caracterizam os regimes de descarga em termos do número de horas por dia, dias por mês ou semanas por ano (”Reg_Desc_Horas_Dia´´, ”Reg_Desc_Dias_Mês´´ e ”Reg_Desc_Semanas´´, respectivamente). Se fossem agregadas, poderíamos ter valores incorrectos, pois ter para uma descarga uma taxa de utilização de 4 semanas por ano, e para outra descarga 3 semanas por ano, não é equivalente a dizer-se que o total são 7 semanas de utilização por ano. Estes períodos podem até ter sido sobrepostos,

Tabela 4.7: Medidas do facto **Desc_AgResid**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Reg_Desc_Horas_Dia	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Average, Max, Min
Reg_Descarg_Dias_Mês	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Average, Max, Min
Reg_Desc_Semanas_Ano	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Average, Max, Min
Caudal_Desc_MedDia	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Sum, Max, Min
Caudal_Desc_MedAno	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Sum, Max, Min
Caudal_Desc_Ponta	-	Instalacao, Geografia, Período, Caract_Descarga, Receptor	Max, Min

pois as descargas podem não ser contínuas. Por este motivo, o operador "Sum" não é aplicado nestas 3 métricas.

No entanto, as duas medidas seguintes podem ter um tratamento diferente. Dado que dizem respeito a valores médios diários e anuais (''Caudal_Desc_MedDia´´ e ''Caudal_Desc_MedAno´´ respectivamente), e são indicadas pelos próprios operadores (motivo pelo qual não têm indicada a fórmula de cálculo), podem efectivamente ser somadas para serem obtidos valores coerentes (um caudal médio anual de descarga adicionado a outro caudal médio anual indica o caudal médio das 2 descargas).

A medida ''Caudal_Desc_Ponta´´ caracteriza o caudal de ponta da descarga, ou seja, o valor instantâneo do caudal da descarga, motivo pelo qual também não é uma medida agregável, a não ser através da utilização dos operadores máximo e mínimo.

De resto, todas as dimensões pelas quais o facto está indexado (à excepção da dimensão **Processo**, conforme referido nos sub-modelos anteriores), podem servir para agregação dos seus registos.

4.4.6 Dados de FontesPontuais

Neste sub-modelo estão incluídos os dados das fontes pontuais e dos equipamentos que contribuem para essas fontes, pois considerou-se que, dada a interligação destes dados, seria relevante analisá-los em conjunto. Por fonte pontual entende-se o ponto de origem das emissões, como por exemplo uma chaminé. Os equipamentos contribuintes são as várias máquinas ou equipamentos que originam as emissões que serão expelidas pelas fontes pontuais.

Na figura 4.18 apresentam-se os dados relativos aos equipamentos contribuintes.

Neste esquema verifica-se que existe uma associação entre três dimensões, a **Instalacao**, a **Fonte_Pontual** e o **Equip_Contrib**. Pretende-se assim representar conceptualmente a relação

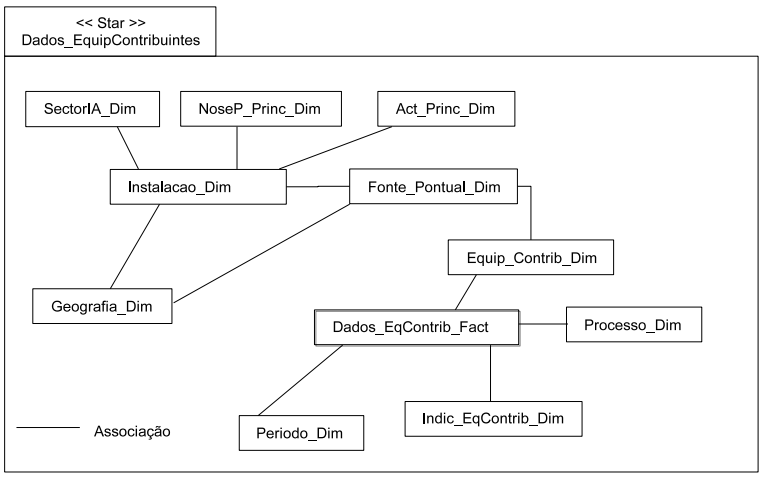

Figura 4.18: Diagrama YAM^2 de nível conceptual mais elevado para os dados relativos a equipamentos contribuintes

entre estas dimensões, pois uma fonte pontual pertence a uma única instalação, e pode ter vários equipamentos contribuintes. A geografia é associada à dimensão **Instalacao** no seu papel caracterizador da instalação, e à dimensão **Fonte_Pontual**, em vez de directamente ao facto, pois assume-se que as emissões por fontes pontuais têm de estar associadas obrigatoriamente a uma fonte, e portanto, a geografia dessa fonte é que será relevante. Actualmente não é recolhida a geografia das fontes pontuais, mas foi já incluída no modelo devido à questão da abrangência.

A opção de implementar os indicadores dos equipamentos contribuintes sob a forma de um dimensão própria (**Indic_EqContrib**), em vez de serem incluídos na dimensão **Equip_Contrib** é justificada pelo facto de ter sido indicado pelo IA que havia mais indicadores aplicáveis a esta entidade, para além dos presentes no formulário PCIP. Assim, colocando a dimensão do indicador à parte, é possível acomodar os vários indicadores, quer sejam provenientes do formulário, quer sejam reportados periodicamente através das fichas de recolha alternativas que o IA recebe dos operadores, com maior frequência que o formulário PCIP.

Alguns dos indicadores relativos aos equipamentos contribuintes são os seguintes:

- Rendimento - Produção de vapor (Kg/h)
- Rendimento - Consumo térmico (MW)
- Consumo máximo de combustível[3] (Kg/h)
- Teor de enxofre (em percentagem) do combustível
- Potencial Calorífico Inferior - PCI do combustível (MJ/Kg)
- Caudal horário (m^3N/h)

Na figura 4.19 apresenta-se o diagrama do esquema conceptual de nível intermédio para os equipamentos contribuintes.

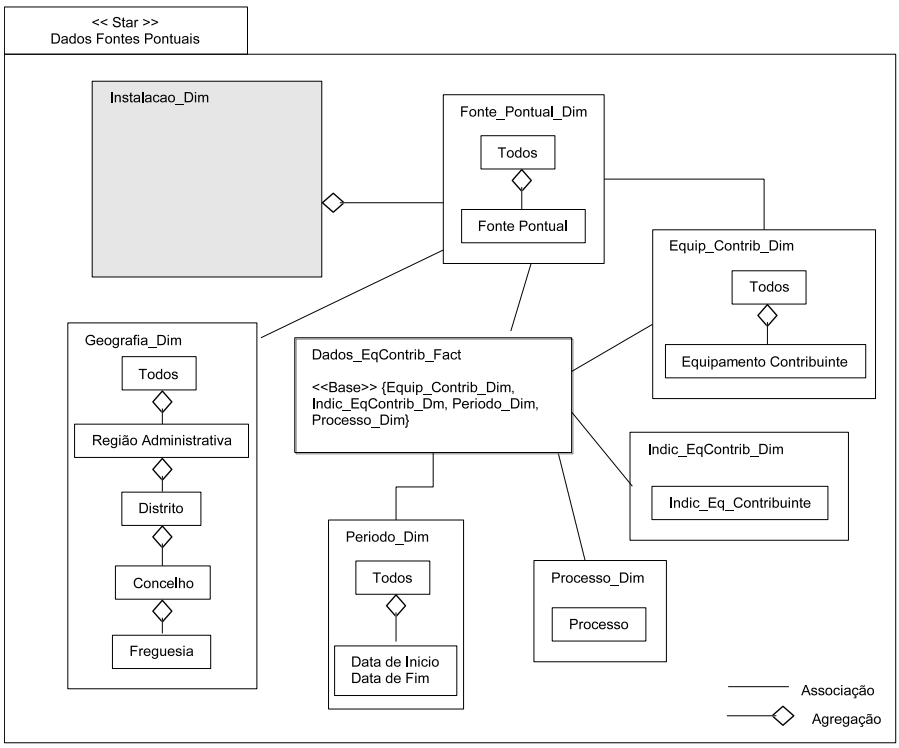

Figura 4.19: Diagrama YAM^2 de nível conceptual intermédio para os dados relativos a equipamentos contribuintes

Na tabela 4.8 apresenta-se a medida do facto **Dados_EqContrib**.

Tabela 4.8: Medida do facto **Dados_EqContrib**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Valor_Indicador	-	Instalacao, Geografia, Período, Fonte_Pontual, Equip_Contrib	-

Este facto irá conter um conjunto de valores caracterizadores destes equipamentos, como por exemplo a produção de vapor, conforme lista de indicadores já apresentada.

Verifica-se que a dimensão **Indic_EqContrib** não faz parte das dimensões de agregação pois, à semelhança dos indicadores de funcionamento e produção, também aqui não faz sentido somar dados relativos a indicadores diferentes. Como já foi descrito na apresentação do esquema conceptual de nível intermédio desta estrela, os indicadores já identificados são de naturezas muito distintas, com unidades e contextos diferentes, portanto não são adicionáveis. Pelo mesmo motivo, as funções de agregação não estão descritas, porque os indicadores podem ou não ser agregáveis consoante a sua natureza (por exemplo, o teor de enxofre (em percentagem) do combustível é à partida um indicador não agregável).

O diagrama de alto nível conceptual para as fontes pontuais é apresentado na figura 4.20.

A motivação da criação da estrela específica para estes dados é que as fontes pontuais têm

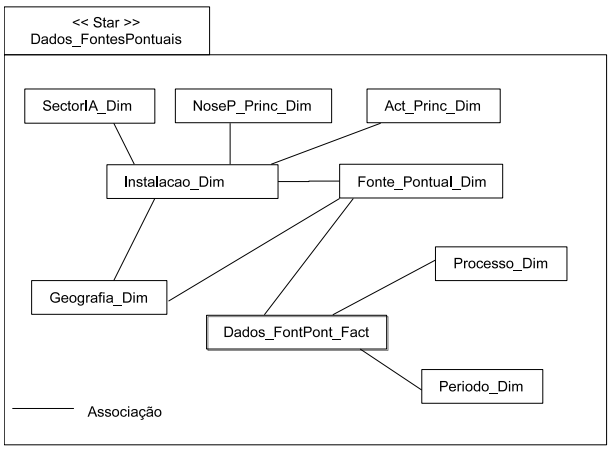

Figura 4.20: Diagrama YAM^2 de nível conceptual mais elevado para os dados relativos a fontes pontuais

efectivamente um conjunto de dados próprios que caracterizam a sua capacidade de emissão, distintos dos dados relativos aos equipamentos contribuintes, e que também são quantificáveis. Assim, este diagrama tem as mesmas dimensões aplicáveis aos equipamentos contribuintes, à excepção das dimensões específicas desses equipamentos, **Indic_EqContrib** e **Equip_Contrib**.

Na figura 4.21 apresenta-se o diagrama de nível conceptual intermédio para as fontes pontuais.

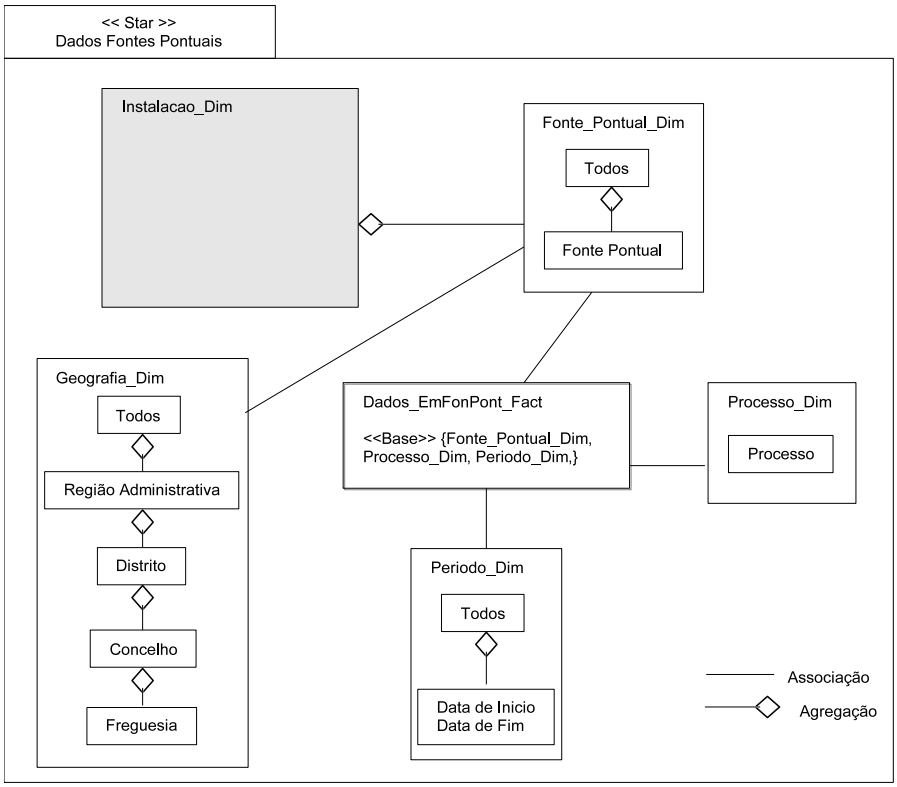

Figura 4.21: Diagrama YAM^2 de nível conceptual intermédio para os dados relativos a fontes pontuais

O diagrama de baixo nível conceptual para o facto **Dados_FontPont** é apresentado na tabela 4.9.

Tabela 4.9: Medidas do facto **Dados_FontPont**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Caudal_ Volum	-	Instalacao, Geografia, Período, Fonte_Pontual	Sum, Average, Max, Min
Velocidade_ Saida	-	Instalacao, Geografia, Período, Fonte_Pontual	Average, Max, Min
Temperatura_ Saida	-	Instalacao, Geografia, Período, Fonte_Pontual	Average, Max, Min

O facto **Dados_FontPont** contém os dados numéricos relativos às emissões expelidas pela fonte pontual, como a velocidade de saída, a temperatura de saída e o caudal.

As medidas "Velocidade_Saida´´ e "Temperatura_Saida´´ apenas podem ser agregados utilizando a média e os operadores máximo e mínimo pois, sendo estas medidas respeitantes a valores instantâneos, considera-se que não faz sentido que sejam somados.

Nesta estrela verifica-se que as medidas são agregáveis por todas as dimensões que lhe estão associadas, à excepção das que não têm associação directa ao facto, e pela dimensão **Processo** já referida.

4.4.7 Dados Socio-Demográficos

Um diagrama que é completamente distinto dos anteriores diz respeito aos indicadores sócio-demográficos. Estes indicadores, embora não sejam recolhidos pelo IA mas sim pelo Instituto Nacional de Estatística (INE), são relevantes para a realização de análises comparativas entre dados do ambiente e os desenvolvimento sócio-demográfico das regiões abrangidas.

A necessidade desta tabela de factos proveio da análise do relatório do estado do ambiente (2003), onde o tipo de análises efectuadas requerem a existência destes dados. Apresentam-se de seguida alguns exemplos de indicadores presentes no REA que utilizam estes dados:

- Evolução relativa das emissões de Gases com Efeito de Estufa (GEE) com o PIB e o consumo de energia primária.
- Comparação entre Portugal e a UE das capitações de GEE, em 2002 - Emissões de GEE per capita em 2002 (toneladas de CO_2 equivalente por habitante).
- Evolução relativa das emissões de substâncias acidificantes com o PIB e consumo de energia primária.
- População residente com sistemas de tratamento de águas residuais.

O diagrama de alto nível conceptual deste esquema é apresentado na figura 4.22.

Neste sub-modelo optou-se por utilizar duas das dimensões utilizadas nos restantes sub-modelos, por uma questão de uniformização: a **Geografia**, e o **Período**. Desta forma, os dados

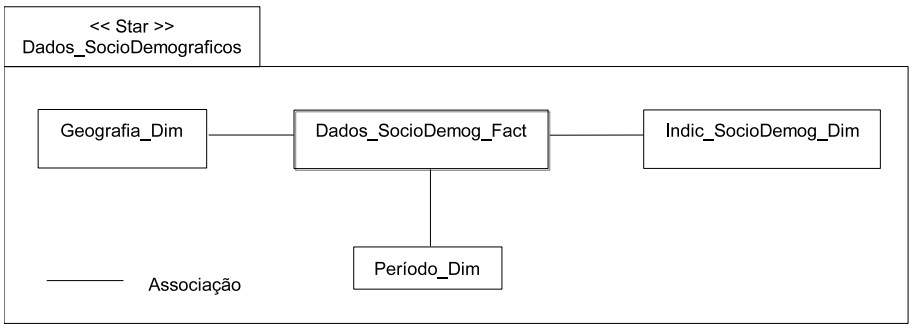

Figura 4.22: Diagrama YAM^2 de nível conceptual superior para os dados demográficos

destes indicadores poderão ser comparados com os restantes dados IA com base nestas dimensões.

Para este efeito, assume-se que os dados fornecidos pelo INE tenham no mínimo uma periodicidade anual e que estejam sempre disponíveis segundo uma área administrativa, como por exemplo o concelho.

Dado o menor nível de detalhe dos dados disponíveis para estes indicadores, o grau de agregação desta tabela é bastante superior ao das outras tabelas, preenchidas a partir dos dados do IA.

Apresenta-se na figura 4.23 o diagrama de nível conceptual intermédio deste esquema.

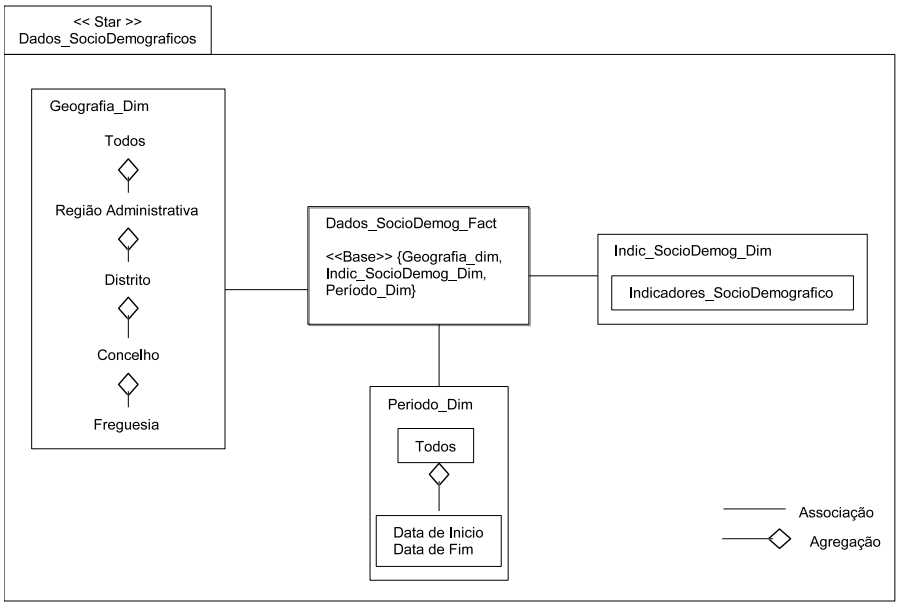

Figura 4.23: Diagrama YAM^2 de nível conceptual intermédio para os dados demográficos

Neste diagrama identifica-se que não se pretende a possibilidade de agregar valores ao longo da dimensão **Período**, porque estes dados não são directamente recolhidos pelo IA, portanto considera-se que a informação disponível sobre o cálculo dos dados não é suficiente para permitir que seja efectuada uma agregação ao longo do período, com garantia de se obterem resultados consistentes.

Também não se pretende agregar ao longo da dimensão **Indic_SocioDemog** devido, mais uma vez, às distintas naturezas de cada indicador.

O diagrama de baixo nível conceptual para o facto **Dados_SocioDemog** apresenta-se de acordo com a tabela 4.10.

Tabela 4.10: Medidas do facto **Dados_SocioDemog**

Nome da medida	Fórmula de cálculo (opcional)	Dimensões de agregação	Funções de agregação
Valor_Indicador	-	Geografia	Sum, Average, Max, Min

Considerou-se que os dados são agregáveis pela dimensão **Geografia**, mas na realidade não há garantias (visto ser outra entidade a efectuar a recolha) de que os dados sócio-demográficos sejam (sempre) recolhidos ao nível mais baixo desta dimensão. Uma abordagem mais cautelosa diria que provavelmente o nível mais baixo a ser considerado para estes dados seria o ”Concelho´´, portanto esta agregação ao longo da **Geografia** tem de ser avaliada e ponderada caso a caso.

4.5 Conclusões

Este modelo conceptual foi corporizado através de 7 sub-modelos, 6 dos quais relativos a dados recolhidos pelo IA, e o sétimo relativo a dados demográficos, cuja relevância foi detectada através da análise do REA de 2003.

Os dados IA utilizados foram recolhidos através de 3 formulários: EPER, PCIP e ficha de recenseamento COV, tendo sido modelizado todo o conteúdo dos formulários EPER e COV. Do formulário PCIP, devido à sua grande extensão, foram incluídas neste modelo apenas algumas componentes consideradas relevantes, seleccionadas com recurso ao conhecimento e opinião dos técnicos do IA. Estes três formulários do IA eram os únicos que estavam já informatizados à data da realização deste trabalho.

O formulário EPER serve de base para os sub-modelos “Dados de emissões de poluentes”, “Dados de produção” e uma parte dos “Dados de funcionamento”. O formulário COV serve de base ao sub-modelo “Dados de COV”. O formulário PCIP recolhe os dados dos restantes sub-modelos à excepção dos dados sócio demográficos(“Dados de descargas de águas residuais”, “Dados de fontes pontuais”), e também para o resto dos indicadores dos “Dados de Funcionamento”.

Para todos estes sub-modelos foram utilizadas dimensões em comum, numa perspectiva de transversalidade, estandardização e maximização dos cruzamentos de dados possíveis.

5

Protótipo

Neste capítulo é apresentado com detalhe o protótipo que foi desenvolvido no âmbito da presente dissertação.

Foi desenvolvido um protótipo consistindo numa base de dados com quatro tabelas de factos, e foi utilizada uma aplicação de construção de relatórios analíticos para uma parte dos dados abrangidos pelo modelo proposto. Neste capítulo será apresentado o âmbito desse protótipo (qual a sua abrangência face ao modelo total), efectuada a descrição da sua implementação e apresentadas as motivações para as várias opções tomadas. É ainda feito um resumo dos resultados obtidos com o protótipo e são apresentadas, na última secção, algumas conclusões.

5.1 Âmbito do protótipo

Para facilitar a descrição do âmbito do protótipo apresentado nesta dissertação, será novamente utilizada a referência à figura 4.1. Tendo como base a estrutura apresentada nesta figura, no protótipo foram considerados os dados relativos à própria instalação, assinalada com o número 1, os **dados de produção**, número 4, e os dados relativos a **emissões de Poluentes para o ar e para a água**, assinaladas com os números 7 e 8 (de notar que estas emissões não estão relacionadas com as fontes pontuais).

Foram seleccionadas estas zonas de dados porque os dados que estavam disponíveis em formato digital, e já de forma estabilizada à data do desenvolvimento do protótipo, diziam respeito ao EPER, que teve uma recolha de dados em 2004 (tinha tido um primeiro exercício em 2002, com carácter experimental). Conforme foi descrito na secção 4.2, o EPER abrange não só dados de emissões de poluentes (para a água e ar) mas também dados relativos a capacidades e volumes de produção.

Assim, e porque o IA sentiu necessidade de ter disponível uma forma rápida que lhe permitisse analisar com fiabilidade e facilidade os dados recebidos, foi desenvolvida esta aplicação protótipo para análise de dados, que contou também com a colaboração dos técnicos do IA para a definição das análises pretendidas.

Os dados EPER incluem os dois tipos de dados de emissões já referidos na secção 4.4.1, ao nível detalhado e ao nível agregado, o que pode ser explicado através da descrição da forma como decorre o processo EPER. É solicitado aos operadores das instalações industriais que relatem, através de um formulário electrónico desenvolvido para o efeito, os valores das suas emissões de poluentes para o ar e para a água, desenvolvidas no âmbito das suas actividades industriais (podem ser PCIP ou não PCIP), e através dos processos industriais (Nose-P) utilizados na instalação. Simultaneamente, solicita-se que relatem alguns dados de funcionamento das instalações (consumos de água e energia), e os volumes de produção, capacidades instaladas e efectivadas para cada uma das actividades da instalação.

Os dados das emissões relatados pelos operadores são depois analisados e verificados pe-

los técnicos do IA, que podem efectuar algumas correcções antes de os dados assumirem um carácter definitivo, e passarem a corporizar os dados de emissões detalhados, denominados "registos IA". Em seguida, é necessário criar os registos de dados de emissões que serão enviados à UE, que estão num nível mais agregado, pois não são detalhados por actividade nem por Nose-P, e sofrem algumas transformações menores, já descritas na secção 4.4.1. Estes dados agregados e transformados designam-se normalmente por "registos UE".

Estava previsto inicialmente serem abrangidos pelo protótipo os dados de funcionamento, mas porque o EPER apenas recolhe um subconjunto de todos os indicadores de funcionamento definidos, os técnicos do IA não consideraram interessante incluí-los.

Desta forma, o EPER é a origem dos dados de três das estrelas que compõem o protótipo, e foram contempladas no modelo conceptual: dados de produção, dados de emissões de poluentes e dados de emissões de poluentes agregadas ao nível da instalação, cuja estrutura se apresenta na figura 5.1.

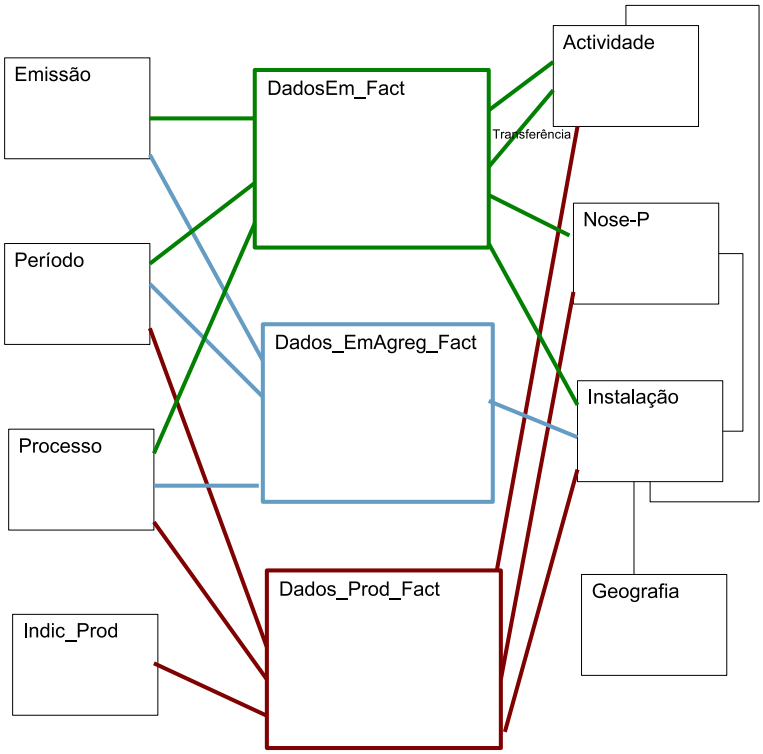

Figura 5.1: Estrutura das estrelas que compõem o protótipo, previstas no modelo conceptual

Como se pode observar, as dimensões aplicáveis à estrela **emissões de poluentes detalhadas** são a **Instalação**, **Actividade**, **Nose-P**, **Emissão**, **Período** e **Processo**. A dupla ligação entre as **Actividades** e o facto **DadosEm** será posteriormente explicada. Para a estrela **emissões de poluentes agregadas** as dimensões aplicáveis são essencialmente as mesmas, com excepção das dimensões **Actividade** e **Nose-P**. Ou seja, todas as emissões de cada poluente, relatadas numa instalação, são agregadas num valor único ao nível da instalação. Para a estrela dos **dados de produção**, as dimensões aplicáveis são **Instalação**, **Actividade**, **Nose-P**, **Período**, **Processo** e

indicadores de produção. Cada uma destas estrelas será detalhada ao longo deste capítulo.

Foi ainda criada uma estrela adicional (não faz parte do modelo conceptual) para cumprir um requisito colocado pelos utilizadores, que pretende facilitar a comparação dos poluentes reportados com a lista de poluentes definida pela UE para as actividades PCIP - para cada actividade PCIP principal, a UE apresenta a lista de poluentes supostos serem emitidos para cada Meio (ar, água) e que devem portanto ser medidos. Esta estrela foi criada com uma dimensão própria denominada **Poluente**, conforme com a dimensão "Emissão"[1] que será detalhada nas **emissões de poluentes**, e pretende cumprir apenas este objectivo específico, para enriquecer a aplicação necessária aos técnicos do IA.

É de salientar que esta implementação num cenário real, embora limitado, permitiu observar várias dificuldades e alguns novos requisitos (um dos quais referido no parágrafo anterior) que poderiam não ter sido levantados se não fosse a necessidade efectiva de utilização por técnicos do IA. Estas questões serão apresentadas ao longo deste capítulo, bem como a representação do esquema final implementado, e será efectuado o seu mapeamento para o modelo conceptual apresentado no capítulo 4.

5.2 Descrição do modelo físico do protótipo

Conforme já foi referido, neste protótipo foram consideradas quatro estrelas distintas, uma das quais é uma estrela auxiliar para a realização de uma análise específica. Para cada uma será apresentado o modelo de dados físico utilizado na sua implementação, efectuando o mapeamento para o modelo conceptual (quando aplicável), e explicadas as opções e diferenças encontradas.

5.2.1 Emissões de Poluentes

Apresenta-se na figura 5.2 o esquema das emissões de poluentes detalhadas. Como podemos observar, este esquema corresponde è estrela das emissões de poluentes detalhadas apresentados no sub-modelo conceptual "Dados das Emissões", tendo no entanto algumas pequenas diferenças de terminologia e associações.

Podemos observar no modelo físico as ligações da **instalação** às **actividades** e aos **Nose-P**, conforme se apresentou no modelo conceptual, distinguindo-se perfeitamente qual a ligação de **actividade** ou **Nose-P** que se está a utilizar: através da actividade principal da **instalação**, ou através das actividades associadas às **emissões** de poluentes.

A dimensão **Geografia** encontra-se apenas associada à instalação pois, conforme já foi referido no modelo conceptual, não existe actualmente a geografia associada directamente às **emissões** de poluentes.

[1] As dimensões conformes são idênticas, ou são subconjuntos matemáticos estrictos das dimensões mais detalhadas [KR02]

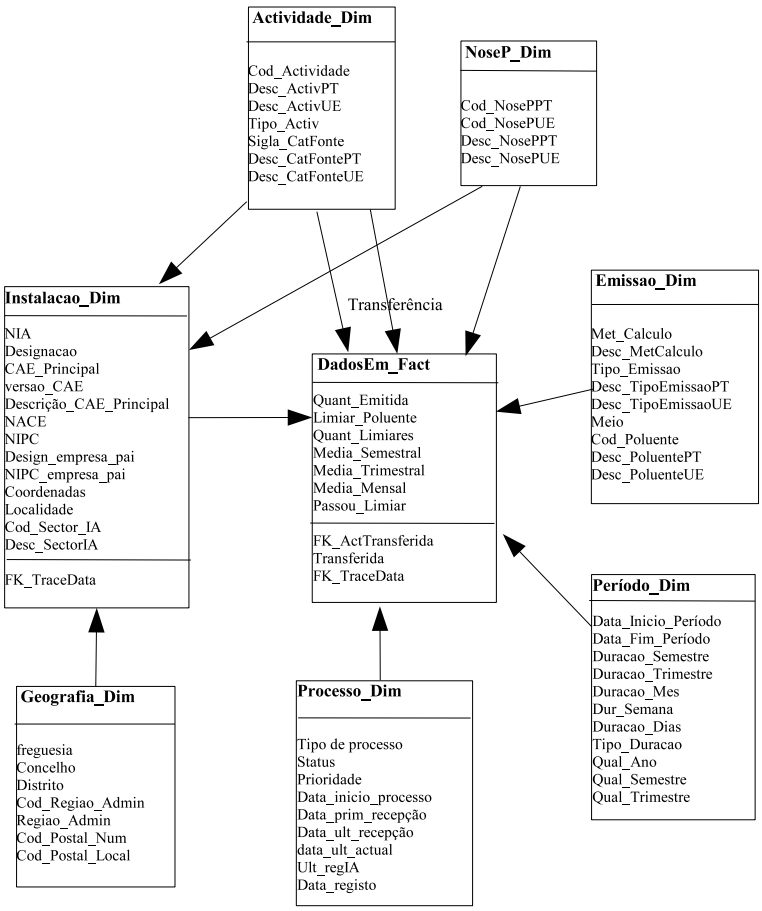

Figura 5.2: Esquema físico das emissões detalhadas

No diagrama físico não foram implementadas as três dimensões **método de cálculo**, **tipo de emissão** e **poluente**, conforme se apresenta no modelo conceptual, tendo as três sido desnormalizadas para uma única tabela de dimensão denominada **Emissão**. Desta forma foi possível criar mais facilmente uma hierarquia de análise entre estas três dimensões, que permite mais facilmente a realização das funções de *drill-up* (alteração da visualização dos dados para um maior nível de agregação) e *drill-down* (alteração da visualização dos dados para um maior nível de detalhe) durante a análise de dados. Nesta hierarquia, criada para efeitos de funcionamento com a aplicação de criação de relatórios, foi definido como nível mais elevado da hierarquia o poluente, sendo que no nível intermédio está o meio (ar ou água), detalhado depois por tipo de emissão (ar, água directo, água indirecto dentro e água indirecto fora). Finalmente, no nível mais baixo da hierarquia ficou o método de cálculo (cálculo, estimativa, medição, indefinido). Desta forma, é possível agregar todos os valores para um determinado tipo de emissão (independentemente da forma de cálculo), depois para um único meio (a água agrega todos os tipos de emissão água directo, água indirecto dentro e água indirecto fora) e, finalmente, para um poluente.

Na sequência da interacção com os técnicos do IA na utilização deste protótipo foi identificado um novo requisito que se pretendia que fosse coberto pela aplicação, respeitante às

emissões detalhadas associadas ao processo nose-P ”109.02´´, e que motivou a segunda associação da tabela de **actividades** à tabela de factos, que não está representada no sub-modelo conceptual das emissões.

Na realidade, este processo Nose-P (bem como as suas subdivisões a 7 dígitos) está directamente relacionado com as Estações de Tratamento de Águas Residuais (ETAR), que do ponto de vista do IA deveriam estar associada à actividade PCIP principal da instalação, pois realizam um processo de “limpeza” dos resíduos provenientes desta actividade. Assim, numa instalação cujo processo Nose-P principal não seja o “109.02” (ou seja, quando a função principal não é a de uma ETAR) que relata emissões associadas ao Nose-P “109.02”, estas devem ser formalmente associadas à actividade principal da instalação, e contabilizadas como tal para efeitos de relato à UE.

Quando se tratem de emissões associadas ao Nose-P ”109.02´´, a associação assinalada na figura 5.2 como “Transferência” referencia a actividade PCIP principal da instalação. Caso contrário, é referenciada a actividade inserida no registo de emissão original. Desta forma, foram preservados os dados originais, e é possível através da instanciação desta nova associação obter os dados das emissões associadas ao Nose-P “109.02” depois de transferidas para a actividade PCIP principal da instalação.

Na tabela de dimensão dos **Processos** é possível observar alguns dos parâmetros que foram recolhidos a partir da informação de processos IA, como por exemplo o estado (*status*), o tipo de processo (COV, EPER, PCIP, etc.) e a data de início do processo.

Na dimensão **Período** foi não só prevista a data de início e fim, como também foi expressa a duração do período a que os dados dizem respeito, em termos de número de semestres, trimestres, meses, semanas e dias, sendo que todos, podem estar vazios, caso ao período em causa não seja aplicável nenhuma destas durações. O período tem também um atributo que o classifica segundo um tipo de duração (semanal, mensal, trimestral), que pode não estar preenchido quando não é aplicável. A qualificação segundo o ano, semestre e trimestre pretendem permitir indexar directamente os períodos respectivos.

Na figura 5.3 apresenta-se o esquema físico das emissões agregadas implementado no protótipo. Este esquema é muito semelhante à estrela apresentada no modelo conceptual para as emissões agregadas, reflectindo apenas as diferenças já descritas na apresentação da estrela das emissões detalhadas de poluentes para o protótipo, no que diz respeito à inclusão das dimensões **método de cálculo**, **tipo de emissão** e **poluente** na mesma tabela de dimensão **Emissão** no protótipo.

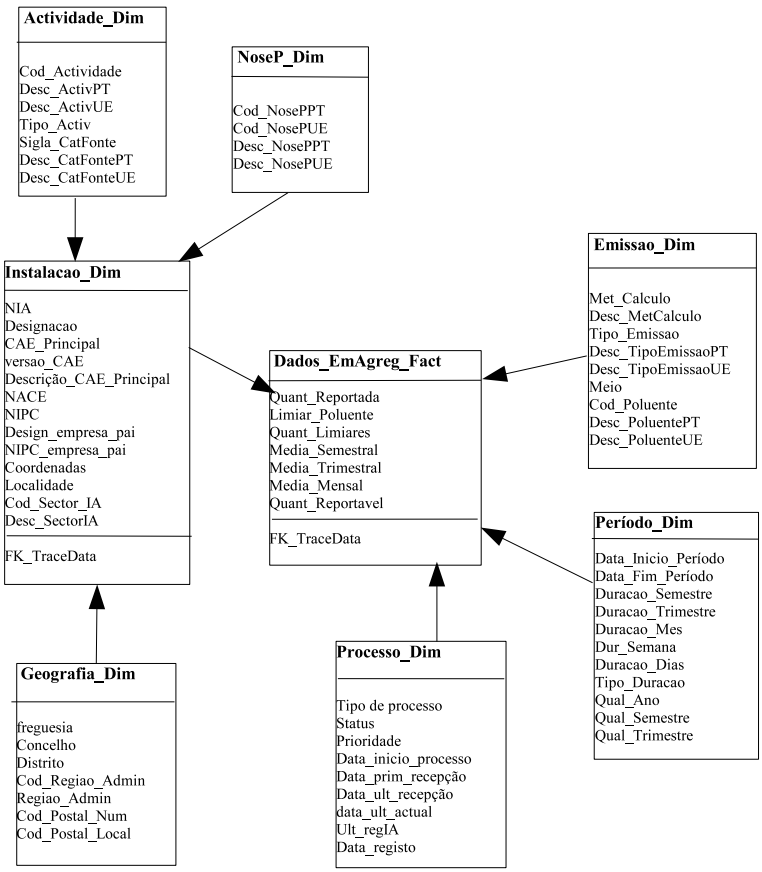

Figura 5.3: Esquema físico das emissões agregadas

Outro requisito (já referido) detectado através da interacção dos técnicos do IA diz respeito à identificação dos poluentes que declaram emissões, por sector IA e meio (ar, água), face à lista de poluentes indexada por sector presentes na documentação de referência para o EPER. Para a implementação deste requisito optou-se pela criação da estrela que se apresenta na figura 5.4.

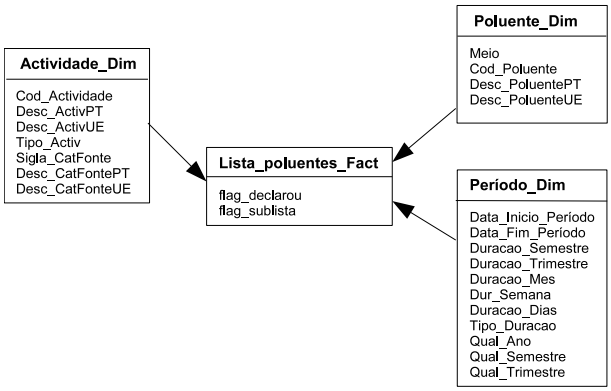

Figura 5.4: Esquema físico para implementação das análises quanto à conformidade da lista de poluentes

Esta é uma estrela agregada, que não utiliza as instalações mas sim as suas actividades principais (associação existente entre a dimensão **Actividade** e a tabela de factos), também designadas "sectores-IA" a que pertence a instalação. A dimensão **Emissão** também não é

utilizada, porque esta análise é revelante apenas para os níveis "meio" e "poluente". Assim, foi criada a dimensão **Poluente** que apenas é utilizada nesta estrela (e que é conforme com a dimensão **Emissão** no sentido indicado em Kimball [KR02]).

Esta estrela é necessária porque, nos dados das emissões, é possível verificar quais os poluentes declarados para cada sector-IA (actividade principal da instalação), meio (ar, água) e poluente, mas não é possível compará-los com o conjunto de poluentes previstos para cada sector pela UE (esta informação não existe nos dados das emissões).

Esta é uma *factless fact table*, ou seja, é uma tabela de factos sem valores numéricos e que apenas foi criada para esta análise específica de comparação de adequação de listas de poluentes por sector-IA. Para a identificação de quais os poluentes que declaram face aos que poderiam declarar foram criadas *flags* em duas colunas separadas, sendo que a associação ao **Período** permite realizar a análise para cada um dos anos onde existem dados. Esta estrela permite saber directamente:

- quais os poluentes que é suposto uma instalação com uma determinada actividade PCIP principal relatar;
- qual a cobertura dos poluentes relatados, considerando todas as instalações de um sector-IA (isto é, com uma determinada actividade PCIP principal);
- e igualmente, embora de forma mais indirecta, comparar os poluentes relatados por uma instalação com a lista de poluentes que eram suposto relatar.

5.2.2 Dados de produção

O esquema físico dos dados de produção, apresentado na figura 5.5 mapeia integralmente o modelo conceptual apresentado na secção 4.4.2.

As tabelas de dimensões **Período**, **Processo**, **Actividade** e **NoseP** foram também já apresentadas no âmbito dos dados de emissões. A dimensão **Indic_Prod** tem apenas dois atributos: um texto, contendo o nome do indicador, e a unidade em que o valor está relatado. A estrutura da tabela de factos é a já apresentada no âmbito do modelo conceptual.

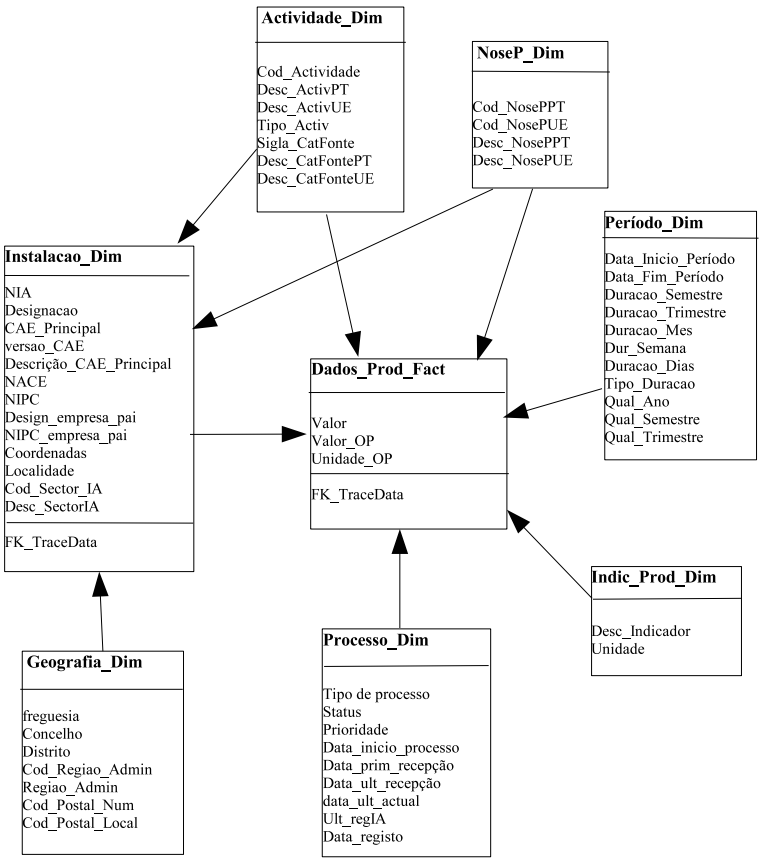

Figura 5.5: Esquema físico dos dados de produção

5.3 Processo de implementação

O protótipo implementado tem os seguintes componentes:

- Base de dados Oracle versão 10G®.
- Oracle Discoverer®, uma aplicação que não tem pré-requisitos específicos de software e de fácil utilização, sendo composta por dois módulos separados: Oracle Discoverer Administrator que permite criar um *End User Layer* onde os metadados serão guardados, e depois permite a criação dos próprios metadados, através da interpretação do esquema físico implementado e com mais alguma informação registada pela pessoa que efectua o desenvolvimento, e o Oracle Discoverer Desktop, uma ferramenta de edição e criação de relatórios.

A motivação de ter sido utilizada a plataforma Oracle, e mais especificamente o Oracle Discoverer®explica-se porque estas eram as ferramentas imediatamente disponíveis e passíveis de serem utilizadas no IA, já com as questões de licenciamento resolvidas. Foram feitas tentativas para fazer a migração da aplicação para o Oracle 10g AS Server®, o que permitiria a utilização do Oracle Discoverer Plus®, com um interface *web* (os utilizadores teriam acesso aos relatórios já criados e à criação de novos relatórios através de um simples *browser*), mas não houve possibilidade em tempo útil de ser criada no IA a infraestrutura necessária em

termos da instalação de software.

A implementação (criação e actualização) do esquema de base de dados de suporte ao protótipo foi efectuada à custa de *scripts* SQL. Para o carregamento dos dados foi utilizado o programa da Oracle SQL Loader®. Os dados foram exportados através de *scrips* SQL a partir da base de dados em ficheiros no formato "comma separated values", e depois transformados para criar as chaves substitutas necessárias para o preenchimento do protótipo. Os dados foram retirados em fases separadas, cada uma correspondente a entidades diferentes:

- Instalações, respectivas empresas e geografia - foram retiradas a partir do módulo transversal ao IA para gestão de entidades(ver figura 1.1 no capítulo 1.1), que inclui um repositório central de dados de entidades. Na sequência desta exportação foi detectada alguma falta de qualidade nos dados destas três entidades. Verificou-se que existem algumas disparidades entre os dados disponíveis nos próprios formulários EPER, onde foram relatadas as emissões, e os dados deste módulo central de gestão de entidades existente no IA. Foi assumido desde o início da implementação deste protótipo que em caso de dúvida os dados que iriam prevalecer seriam os do repositório central (de onde foram efectivamente exportados), mas nalguns casos foi mesmo detectada a necessidade de correcção do conteúdo do repositório central.

- Dados de emissões provenientes dos registos IA e UE- retirados a partir da base de dados operacional de suporte aos formulários (Gestão dos E-forms)

- Dados de processo - para permitir identificar quais as prioridades de processo associadas a cada instalação, retirados também a partir da base de dados operacional de suporte aos formulários (Gestão dos E-forms).

- Dados de volumes e capacidades - foram exportados a partir da base de dados operacional de suporte aos formulários, tendo sido acrescentada em cada linha a indicação de qual o indicador associado a essa linha.

As listas classificativas, **actividades PCIP**, **Nose-P**, **indicadores de produção**, **processo** e **período** (apenas existiam dois períodos de dados) foram definidas como tabelas estáticas, pois são listas definidas e limitadas. A lista de actividades PCIP e Nose-P foi fornecida pelo IA, pois faz parte da informação de apoio ao próprio formulário EPER.

Para as **actividades não-PCIP**, que no formulário EPER foram corporizadas como simples campos textuais de inserção livre, e portanto pouco relevantes para análise, foi tomada a opção, após análise conjunto com os técnicos do IA, de se substituir a sua descrição (às vezes contendo mais de 1000 caracteres, e portanto até tornava impraticável a leitura dos relatórios) pelo respectivo Nose-P da emissão. Esta opção, embora tivesse mostrado os seus frutos em temos de análise de dados, teve também os seus inconvenientes, pois verificou-se que existiam casos

em que para um mesmo processo nose-p havia várias actividades não PCIP com emissões associadas. Foram apresentadas três propostas de correcção desta problemática aos técnicos IA, nomeadamente:

1. Ser criada uma lista finita, delimitada e abrangente de caracterizações para as actividades não PCIP, e em cada registo de emissão associado a uma actividade não PCIP esta ser substituída por um elemento da lista.
2. Os valores associados a um mesmo processo Nose-P com várias actividades não PCIP diferentes serem somados, de forma a criar um único registo.
3. Serem criados novos valores para as descrições das actividades não PCIP em que houvesse conflito (depois de terem sido substituídas pelo processo nose-P), para que um registo ficasse associado à actividade original, e o outro ficasse associado ao novo valor.

A primeira hipótese foi considerada pelos técnicos do IA como não sendo praticável, devido ao volume de dados e de alterações envolvido. Quanto à segunda hipótese, tinha a desvantagem de poder adulterar as contagens de emissões, pois haveria casos em que 2 ou mesmo 3 registos passariam a contar apenas como uma única emissão, de maior valor. Assim, a resolução adoptada foi a indicada no ponto 3, em que à descrição original da actividade se acrescentou o sufixo "-v2´´ para diferenciar valores (por exemplo, para a actividade não PCIP já transformada "105.01´´ foi criada uma nova actividade "105.01-v2´´).

A dimensão **emissão** foi preenchida através da selecção de todos os valores diferentes de poluente, tipo de emissão e método de cálculo presentes nos registos de emissões exportados.

5.4 Resultados obtidos

O resultado final deste protótipo foi corporizado por um conjunto de relatórios interactivos implementados sobre o Discoverer, definidos em conjunto com os técnicos do IA, e que serviram para a realização das análises dos dados das emissões relatadas no âmbito do EPER 2004 (embora tenham sido igualmente tidos em consideração dados do EPER 2002).

Cada um dos relatórios construídos sobre o Discoverer funciona de forma interactiva, isto é, permite nomeadamente que os utilizadores alterem a forma de visualização, o detalhe com que os resultados são apresentados e que apliquem diferentes filtros sobre os dados envolvidos na análise pretendida. A interface que se apresenta ao utilizador é a típica usada em ferramentas OLAP. Na figura 5.6 apresenta-se um exemplo da interface disponível, num relatório que apresenta a distibuição do número de instalações que relataram os vários poluentes, para o ano de 2002, meio "ar" e sector-IA 2.6 ("Instalações de tratamento de superfície de metais e matérias plásticas que utilizem um processo electrolítico ou químico, quando o volume das cubas utilizadas nos banhos de tratamento realizado for superior a 30 m^3").

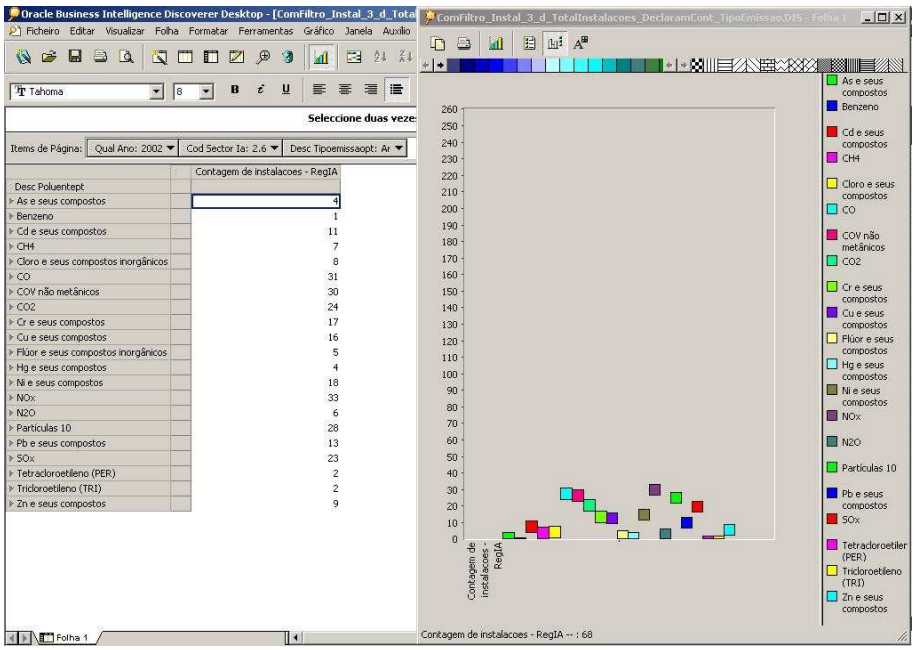

Figura 5.6: Análise do número de instalações que relatam cada poluente

As análises desenvolvidas podem ser agrupadas do seguinte modo:

- Análises sobre quais e quantas as instalações que participaram, isto é: como se distribuem geograficamente pelo País, como se distribuem por actividade PCIP principal, Nose-p principal ou Código de Actividade Económica; quais e quantas as que reportaram apenas em 2002, apenas em 2004 e em ambos os anos.

- Análises do padrão industrial das instalações envolvidas, isto é, caracterização do conjunto de actividades PCIP e dos Nose-P das emissões declaradas. Com esta análise pretende-se verificar quais as relações entre actividades e processos industriais relatadas pelas instalações, para validar a sua conformidade com o esperado (listas de relações entre as actividades PCIP e os Nose-P definidas pela UE), tendo em conta a actividade principal de cada instalação.

- Avaliação dos padrões de declaração de emissões: ordenação das instalações por quantidade de poluentes relatados; evolução das quantidades de poluentes relatados em 2004 face a 2002; verificação das instalações que não relatam dados à UE, mas que para determinados poluentes ficam perto do limiar do poluente; poluentes declarados e relatados por instalação, Nose-P e actividades das emissões. Naturalmente que são possíveis (e usados) cruzamentos destas distribuições e contagens, como por exemplo os Nose-P referentes apenas às instalações cuja actividade PCIP principal seja a 1.1. ("Instalações de combustão com potência calorífica de combustão superior a 50 MW").

- Verificação da lista de poluentes declarados face à lista de poluentes da UE, por actividade principal (sector-IA) da instalação. Esta verificação tem como objectivo o controlo

de qualidade da lista de poluentes definidos pela UE: se uma instalação não abranger todos os poluentes, poderá estar-se numa situação de relato insuficiente de emissões, ou que a lista não esteja adaptada à realidade do sector-IA da instalação em causa; se uma instalação abranger mais poluentes do que os definidos na lista, poderá ser uma situação de incompletude da lista de poluentes para o sector-IA da instalação, e portanto precisará de ser revista.

- Análise das quantidades emitidas tendo em conta as transferências das emissões associadas às ETAR para a actividade principal das instalações. Esta análise tem dois objectivos: verificar a conformidade dos valores relatados à UE, que devem utilizar as emissões das ETAR associadas à actividade principal da instalação; avaliar o impacto das ETAR ao nível da taxa de relato de emissões à UE.

- Análise dos valores de volumes de produção, capacidade instalada e capacidade efectivada; taxa de eficiência das instalações ("quantidades emitidas"/ "volume de produção"). Pretende-se com esta análise verificar quais as instalações que retiram maior rentabilidade da sua actividade, com os menores valores possíveis de emissões. Uma instalação pode ter grandes valores de emissões por um de dois motivos: é efectivamente uma grande instalação em termos de produção, pelo que a sua taxa de eficiência é alta; ou é uma instalação de dimensão relativamente pequena, mas que não utiliza processos industriais que minimizem as emissões.

- Análises sectoriais - este conjunto de análises inclui algumas das já apresentadas, mas detalhadas ao nível de cada sector-IA. Pretende-se com estas análises verificar a conformidade da classificação por sectores definida pelo IA (os sectores correspondem às actividades principais das instalações), e a divisão de cada sector nos seus subsectores (por exempplo, um subsector possível será o tipo de combustível utilizado pelas instalações do sector) sempre que aplicável.

O gráfico apresentado na figura 5.7 representa a evolução temporal das emissões declaradas do poluente "Pb e seus compostos", para os vários meios, para os anos 2002 e 2004.

Para vários relatórios foi utilizada a possibilidade de exportação para EXCEL®(é possível observar na figura 5.7 na barra de ferramentas superior, do lado direito, o botão para exportação), permitindo criar maior variedade de gráficos e aproveitando as potencialidades desta ferramenta para o manuseamento de dados, como por exemplo a utilização de tabelas *pivot*. Apresenta-se na figura 5.8 um exemplo de um gráfico criado desta forma.

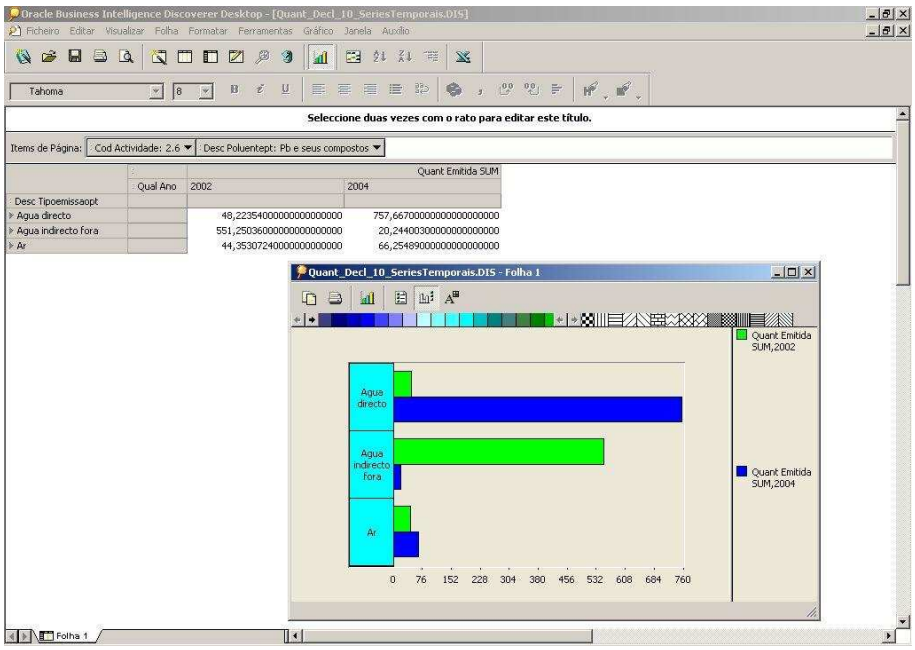

Figura 5.7: Evolução 2002-2004 dos dados de emissões relatados

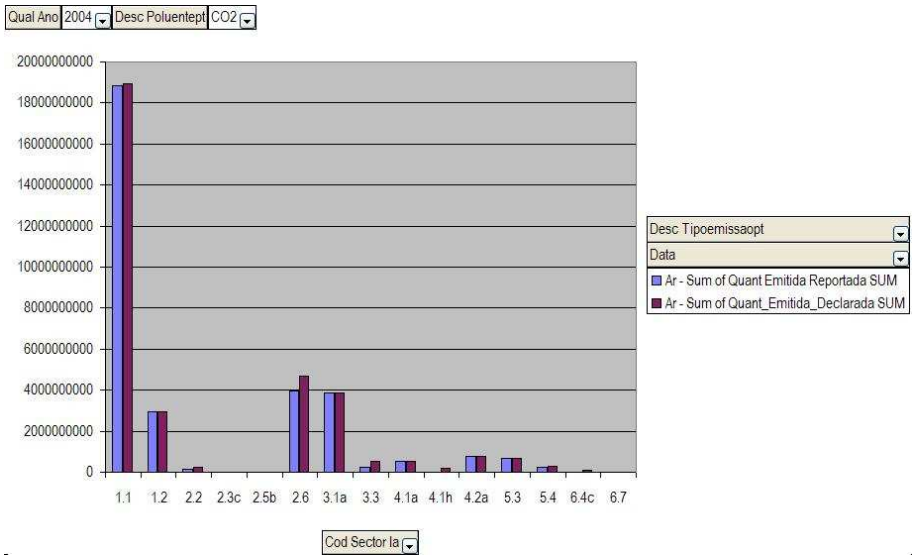

Figura 5.8: Emissões relatadas e emitidas para o CO2, em 2004

Neste último gráfico apresenta-se para o ano 2004, poluente CO2, a quantidade emitida face à quantidade relatada à UE, para os vários sectores-IA. Se existirem grandes diferenças entre estes dois valores, significa que existem várias instalações com pequenas emissões, que não alcançam os limiares do poluente, mas que somadas alcançam um valor visível. Se esta situação ocorresse com muita frequência, isso poderia implicar que os limiares deste poluente estariam demasiado elevados (análise que foi realizada e cujo resumo se apresenta como anexo na secção 7.1).

No anexo 7.2 apresenta-se a a composição de cada relatório em termos de métricas e eixos de análise, sob a forma matricial, sendo que as letras apresentadas na matriz indicam a posição

do eixos de análise no relatório: L=linha, C=coluna, P=página.

5.5 Conclusões

Este protótipo teve uma boa aceitação por parte dos utilizadores pois permitiu:

- Controlar a qualidade dos dados, quer durante a fase de carregamento, onde foram efectuadas as verificações dos dados das instalações, respectivas empresas e geografia já referidas, quer posteriormente através da própria análise dos resultados.
- Verificar a completude do universo das instalações que declaram poluentes, face ao número expectável de instalações que deveriam ter declarado.
- Caracterizar o padrão das emissões de poluentes para cada um dos sectores-IA, tendo em conta os Nose-P utilizados pelas instalações presentes em cada sector.
- Iniciar um trabalho de determinação de índices de poluição, tendo em conta não só os valores das emissões como também os volumes de produção e as capacidades (instalada e efectivada) das instalações.
- Comparar as declarações de emissões de cada instalação à luz dos padrões de emissões de instalações no mesmo sector IA.

A ferramenta utilizada demonstrou ter algumas limitações na sua utilização, principalmente nas àreas da visualização dos resultados e na realização de cruzamentos entre tabelas de factos diferentes. A utilização de outra ferramenta poderia ter facilitado a utilização do protótipo por parte dos utilizadores, que sentiram algumas dificuldades na sua adaptação à ferramenta, e potenciado uma maior exploração da informação disponível.

6

Conclusões e trabalho futuro

Neste capítulo apresenta-se um resumo do trabalho realizado ao longo desta tese, e tecem-se algumas conclusões sobre o desenvolvimento de SADA. São ainda deixadas aqui algumas sugestões de trabalho futuro.

Este é o último capítulo desta dissertação, onde se pretende apresentar a compilação das conclusões que foram retiradas no âmbito deste trabalho, e se apresentam algumas sugestões, baseadas no estudo efectuado e no que foi aprendido, sobre possíveis direcções a serem seguidas para trabalho futuro.

6.1 Resumo

No âmbito deste trabalho foram realizadas várias actividades, nomeadamente:

- Realização de uma análise da conformidade dos limiares de poluentes definidos pela UE relativamente à realidade portuguesa, cujas conclusões são resumidas no anexo 7.1.
- Estudo de trabalhos e desenvolvimentos relacionados com a área de SADA. Foram analisados e resumidos vários sistemas relacionados com esta área: o Le Select, o projecto SIMAGE, o sistema da instalação Pantex e o Envirofacts.
- Pesquisa de propostas de modelação conceptual que pudessem ser utilizadas para apresentar os modelos definidos no âmbito desta dissertação. Foram analisados várias propostas, mas que se verificou não serem reconhecidas a nível abrangente, tendo sido utilizadas apenas em alguns projectos de âmbito específico. Foi seleccionado o YAM^2, que após algumas adaptações foi utilizado para a apresentação do modelo conceptual resultante desta dissertação.
- Implementação de um protótipo de âmbito reduzido, mas que foi efectivamente utilizado pelos técnicos do IA como ferramenta de análise e verificação de dados. Esta implementação abrangeu as tarefas de controlo de qualidade dos dados, definição do ETL para povoamento da aplicação, implementação da base de dados de suporte, definição da estrutura de metadados associada à utilização da ferramenta e apoio aos técnicos do IA na sua utilização.

6.2 Conclusões

Foi possível verificar, através do estudo realizado e do protótipo implementado, que as problemáticas normalmente associadas ao desenho e implementação de um DW têm especial impacto nos SADA. Na realidade, as questões relacionadas com controlo de qualidade, rastreabilidade, SCDs e dimensão tempo ganham outra complexidade quando se está a abordar um tema tão abrangente, complexo e diversificado como é o ambiente. A complexidade deste tema resulta de vários factores:

- Interdisciplinaridade - Para se ter uma visão abrangente sobre a situação ambiental é necessário ter em conta os 3 eixos já referidos: ambiental, social, e económico. Assim,

existem vários tipos de análises e relacionamentos que podem ser definidos para uma abordagem transversal à temática ambiental, abrangendo várias disciplinas e áreas de estudo.

- Difícil recolha e controlo de qualidade dos dados - desde a recolha da informação ao seu tratamento e posterior utilização estão envolvidos diversos actores, sendo que as fontes de informação ambiental são muitas e de carácter variado. Os maiores colectores de informação ambiental continuam a ser as entidades governamentais, que estão dependentes da informação fornecida pelas indústrias, ou recolhida através de redes de monitorização e estudos efectuados.

- Evolução dos modelos e interesses ao longo do tempo - os indicadores ambientais estão permanentemente em actualização à medida que a temática ambiental é estudada com maior detalhe e surgem novas necessidades. Por exemplo, o formulário EPER retractado neste trabalho será substituído a partir de 2007 pelo formulário PRTR (Registo Europeu das Emissões e Transferências de Poluentes), uma iniciativa de âmbito mundial, sob a alçada da ONU. Estes dois formulários terão conteúdos parecidos, mas o PRTR será efectivamente uma extensão do EPER (até porque será uma iniciativa mundial, por contraposição à iniciativa europeia do EPER), incluindo também a área dos resíduos e abrangendo um maior número de poluentes do que os anteriormente cobertos pelo EPER.

- Métricas/aditividade - as métricas envolvidas no SADA são de diversas naturezas, e abrangem várias temáticas distintas, o que se justifica também pela interdisciplinaridade envolvida nesta temática, conforme já descrito. Assim, as características de aditividade, semi-aditividade ou não-aditividade adquirem especial importância neste contexto, merecendo uma análise mais detalhada.

- Necessidade de uma abordagem global - quando falamos de ambiente não podemos apenas falar de determinadas zonas ou regiões pois, conforme se demonstrou ao longo desta dissertação, há uma necessidade efectiva de abordar o ambiente a nível global. Afinal de contas, todos os ecossistemas existentes neste planeta estão relacionados, pelo que não podem ser apenas considerados como subconjuntos isolados e limitados.

Do estudo efectuado, não foi possível perceber que existissem muitas aplicações SADA de grande dimensão suficientente detalhadas e estudadas para servir de apoio efectivo ao âmbito deste trabalho. Assim, considerou-se que este não é um domínio no qual haja muita tradição do uso das técnicas de OLAP e DW. Os SAD aplicados à área ambiental têm sido principalmente focados na utilização de modelos, e só recentemente se está a adoptar os SAD baseados em dados, em grande parte devido à quantidade de informação que cada vez mais é necessária para a realização de análises sobre o ambiente.

No caso específico de Portugal a situação é um pouco mais grave, pois existe pouca tradição no relato de dados reais e nos processos de certificação de qualidade, ainda menos nos processos de certificação de qualidade ambiental. A camada industrial portuguesa é constituída maioritariamente por pequenas e médias empresas (por exemplo, lavandarias, suiniculturas), cujo nível médio de educação dificulta o relato de dados e de informação e a sua qualidade, principalmente quando os temas envolvidos são complexos e ainda pouco treinados, como é o caso da informação ambiental. Mesmo ao nível do estado os inventários informatizados são ainda poucos e recentes, como por exemplo o caso do EPER. O registo de emissões é recente em Portugal, mas já existe noutros países europeus há vários anos.

Ao nível dos modelos conceptuais, verificou-se uma lacuna no que diz respeito a modelos efectivamente usados e reconhecidos de uma forma abrangente (os modelos encontrados diziam respeito a projectos protótipo, tipicamente de carácter académico). Mesmo o CWM, a proposta de *standard* iniciada pelo grupo OMG, tem estado aparentemente inactivo desde 2004, data em que foi publicado o último documento disponível na internet. O modelo YAM^2 foi baseado neste *standard*, que após algumas adaptações se verificou ser um modelo que permitiria cumprir os requisitos de expressividade que se pretendiam obter neste trabalho.

O protótipo foi desenvolvido em ambiente real, com os contributos dos utilizadores e de uma forma iterativa. Verificou-se que este método de trabalho foi essencial para o sucesso do desenvolvimento desta aplicação e que portanto se considera ser uma boa prática no desenvolvimento destes sistemas, principalmente quando envolvem áreas tão específicas como é o caso do Ambiente. A aplicação foi efectivamente utilizada pelos técnicos do IA, tendo tido um nível aceitável de acolhimento por parte dos utilizadores.

6.3 Trabalho futuro

Podem-se referir várias direcções possíveis para trabalho futuro, nomeadamente:

- Possibilidade de abordar com maior detalhe a questão dos processos ETL, que foram apenas referidos de forma superficial no âmbito do protótipo implementado.
- Estender e melhorar este protótipo, para que abranja mais áreas de dados do IA e aumentando as suas potencialidades, utilizando outras ferramentas e instrumentos (por exemplo, num ambiente *Web*), ou mesmo através da integração com SIG (conforme já demonstrado na tese "Integração de Informação Geográfica em Sistemas OLAP", efectuada pela Dra. Rosa Matias [Mat06], esta seria uma abordagem possível e interessante).
- Estender o modelo conceptual para que abranja dados de gestão e, ao nível ambiental, os dados relativos ao comércio de emissões, um processo que se encontra actualmente a ser implantado no IA. Ao nível dos processos já estabilizados e possíveis candidatos a serem

incorporados na extensão destaca-se o SEVESO e os processos AIA.

- Sectorizar as instalações, de forma a permitir uma melhor avaliação dos dados relatados, quer através da utilização de técnicas *Data Mining*, quer através de modelos para análises específicas.
- Ponderar a longo prazo a possibilidade de se criar uma plataforma de mediação e publicação de dados do ambiente em Portugal. Para isto seria necessário definir uma taxonomia [1] de dados do ambiente, e efectuar um estudo tecnológico das possibilidades de implementação desta plataforma.
- Estender o DWA/SADA de forma a incorporar também informação sobre o estado do ambiente.

[1] Classificação de coisas ou os princípios subjacentes da classificação.

7

Anexos

Anexos da dissertação.

7.1 Conclusões do relatório de análise dos limiares

Foi realizada em 2004 uma análise dos limiares, cujo resultado foi transposto para o relatório interno "Análise dos dados dos registos europeus - EPER 2002" [LP04]. Nesta análise foram utilizadas algumas métricas, cujo significado se apresenta a seguir:

- TRE - a Taxa de *Reporting* (relativa às Emissões) representa a percentagem de quantidades emitidas e relatadas à UE. Esta taxa é obtida pela seguinte fórmula: "Quantidade de emissões que foram relatadas, considerando todas as instalações que emitiram esse poluente em quantidades superiores ao limiar definido pela UE para esse poluente e meio" / "Quantidade total de emissões registada no IA para esse poluente e meio, considerando todas as instalações que emitiram esse poluente". Pretende-se que: TRE >= 90%.
- P - A Permissividade serve para verificar se os limiares estão a cumprir o seu objectivo de diminuir o número de instalações envolvidas no relato de dados à UE, maximizando as emissões relatadas, pois transcreve a permissividade dos actuais limiares, e permite avaliar qual a redução obtida sobre o número de instalações que relatam. Quanto menor for o valor deste rácio, melhor será a sua eficácia, desde que os valores de TRE permaneçam o mais elevados possiveis. A permissividade é obtida através da formula "Número de instalações registadas que relataram emissões à UE para determinado poluente e meio, ou seja, emitiram valores para esse poluente e meio superiores aos limiares definidos pela UE" / ("Número de instalações registadas no IA que apresentaram emissões não nulas para determinado poluente e meio" - "Número de instalações registadas que relataram emissões à UE para determinado poluente e meio, ou seja, emitiram valores para esse poluente e meio superiores aos limiares definidos pela UE"). Pretende-se que: P <= 80%, situação em que menos de metade das instalações que emitiram também relataram à UE

Do estudo efectuado, foi possível concluir que de um modo geral os valores dos limiares de poluentes definidos pela UE se encontram razoavelmente adaptados à realidade portuguesa, já que dos 63 poluentes emitidos em Portugal (26 para água e 37 para o ar) apenas 11 (3 para a água e 8 para o ar) se consideraram como precisando de uma análise cuidada em termos de valores dos limiares, porque o volume de emissões relatado à UE não atinge os 90% do total de emissões. Para todos os outros, ou são atingidos os valores de relato desejados ao nível da UE (pelo menos 90% das emissões relatadas), ou o valor total emitido é bastante pequeno, na ordem de um limiar ou menos.

Da análise detalhada efectuada a estes 11 poluentes, concluiu-se o seguinte:

- Para um deles (NH3 / Ar) não se considera compensatória a diminuição do valor do limiar devido ao grande aumento da permissividade que se verifica.

- Para seis dos poluentes deverá ser seriamente tomada em consideração a diminuição dos valores dos limiares, segundo a tabela que se apresenta a seguir:

Tabela 7.1: Poluentes para os quais poderá ser tomada em consideração a diminuição dos valores dos limiares

Poluente / Meio	Redução proposta	Aumento TRE	Valor de TRE atingido	Aumento P	Valor de P atingido	Observações
Cianetos / Agua	1/5	40,27	79,08	8,8	11,43	
COV não metânicos / Ar	2/5	11,12	88,82	7,51	12,64	Possibilidade de 1/5, mas com maior aumento de permissividade
Flúor e seus compostos inorgânicos / Ar	3/5	5,34	89,18	9,14	16,95	Possibilidade de 2/5 mas com maior aumento de permissividade
Fósforo - total / Água	2/5	11,04	85,46	13,55	23,77	Possibilidade de 1/5, mas com maior aumento de permissividade
Hg e seus compostos / Ar	2/5	9,2	89,66	10,35	24,24	Possibilidade de 1/5, mas com maior aumento de permissividade
PCDD+PCDF (dioxinas + furanos) / Ar	4/5	17,76	94,28	4,97	9,52	

- Quanto aos pares (Poluente, Meio): Cloro e seus compostos inorgânicos / Ar, Pb e seus compostos / Ar, PM10 / Ar, os seus valores de TRE com os limiares actuais são relativamente perto dos 90%, e portanto deverá ser ponderado se compensa a diminuição dos valores dos limiares.

- Quanto ao último poluente (Azoto - Total / Água), verifica-se que nenhum dos intervalos de limiar considerados nesta análise consegue fazer com que seja atingido o objectivo de reporting (90%) para este poluente. Mesmo a diminuição do limiar para 1/5 do actual apenas consegue atingir uma TRE de 85%. No entanto, o aumento da permissividade (20%, para um valor total de 26%) quando se reduz o valor do limiar para 1/5 é compensado pelo ganho em *reporting* (25%), embora não largamente. Deve assim ser ponderada

a relação custo-beneficio da alteração do limiar deste poluente.

7.2 Matriz de composição dos relatórios criados

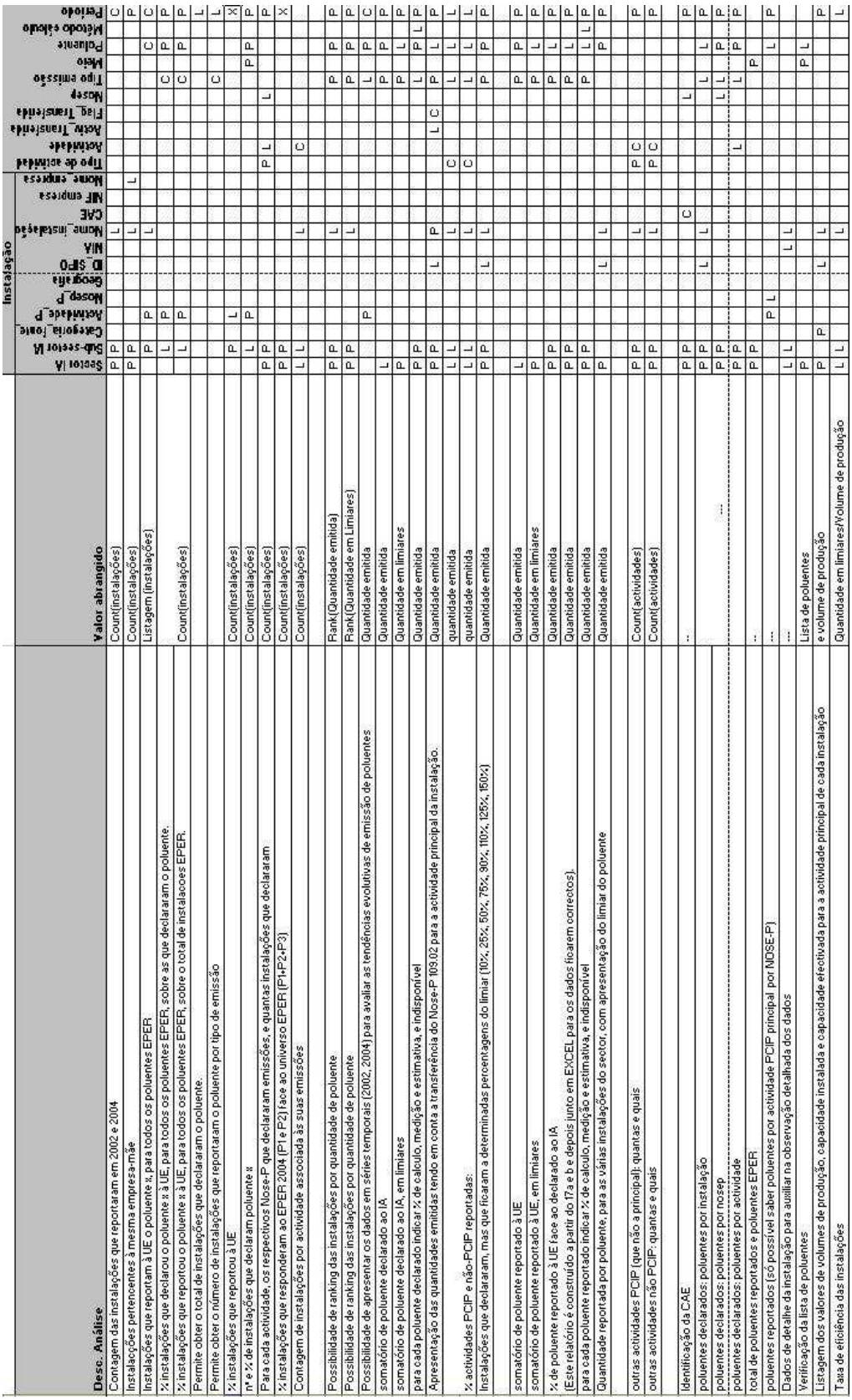

Figura 7.1: Matriz de relatórios

Bibliografia

[1ke] 1keydata.com. Bill inmon vs. ralph kimball. http://www.1keydata.com/datawarehousing/inmon-kimball.html.

[Agea] Environment Agency. Environmental indicators. http://www.environment-agency.gov.uk/yourenv/432430/432593/?version=1&lang=-e.

[Ageb] US Environmental Protection Agency. Envirofacts data warehouse. http://www.epa.gov/enviro/html/ef_overview.html.

[Ass] Ralph Kimball Associates. http://www.kimballgroup.com/.

[ASS02] Alberto Abelló, José Samos, and Fèlix Saltor. Yam^2 (yet another multidimensional model). *http://www.lsi.upc.es/aabello/publications/home.html*, 2002.

[Bat] Ion Leon Batachia. Towards user-centred and cost-effective development of environmental decision support systems. http://www.ici.ro/ici/revista/sic1999_4/art03.html.

[BT] Inc BWX Technologies. Powering transformation. http://www.bwxt.com/.

[CCC+01] IBM Corporation, Unisys Corporation, NCR Corporation, Hyperion Solutions, Oracle Corporation, UBS AG, Genesis Development Corporation, and Dimension EDI. *Common Warehouse Metamodel (CWM) Specification*. Object Management Group (OMG), 2001.

[Cec01] Luigi Ceccaroni. Ontowedss - an ontology-based environmental decision-support system for the management of wastewater treatment plants. *http://www.angelfire.com/scifi2/technopapa/thesis.pdf*, 2001.

[Cena] RFP Evaluation Centers. Analytical hierarchy process (ahp). http://www.rfp-templates.com/Analytical-Hierarchy-Process-(AHP).html.

[Cenb] RFP Evaluation Centers. Multi-criteria decision-making (mcdm). http://www.rfp-templates.com/Multi-Criteria-Decision-Making-MCDM.html.

[Cona] W3C World Wide Web Consortium. Extensible markup language (xml). http://www.w3.org/XML/.

[Conb] W3C World Wide Web Consortium. Hypertext markup language (html) home page. http://www.w3.org/MarkUp/.

[Conc] W3C World Wide Web Consortium. Overview of sgml resources. http://www.w3.org/MarkUp/SGML/.

[Cou] Indiana Geographic Information Council. How are people using gis? http://www.in.gov/igic/realworld/index.html.

[CW00] Yingwei Cui and Jennifer Widom. Lineage tracing in a datawarehousing system. *http://csdl.computer.org/comp/proceedings/icde/2000/0506/00/05060683.pdf*, 2000.

[dA04] Instituto do Ambiente. Instituto do ambiente. www.iambiente.pt, 2004. Site do Instituto do Ambiente.

[dAeOdT04] Inspecção-Geral do Ambiente e Ordenamento do Território. Temática das tintas e vernizes. http://www.ig-amb.pt/documentos/RelatorioTematico/RT-TematicaTintaseVernizes.pdf, 2004.

[dCNRPB03] Miguel de Castro Neto, Maria Brandão L. Rodrigues, Pedro Aguiar Pinto, and Isabel Berger. Traceability on the web - a prototype for the portuguese beef sector. *http://www.date.hu/efita2003/centre/pdf/1106.pdf*, Julho 2003.

[ddAICdT99] Centre d'Estudis d'Informació Ambiental Institut Catalá de Tecnologia. A new model of environmental communication for europe from consumption to use of information. Technical report, European Environment Agency(EEA), 1999.

[dR03] Diário da República. Decreto-lei 113/2003, de 4 de junho. Diário Republica SÉRIE I-A, número 129, de 4 de Junho de 2003, Junho 2003.

[EEA02] Environmental signals 2002 - benchmarking the millennium (environmental assessment report no 9). Technical report, EEA (European Environment Agency), 2002.

[(EP] US Environmental Protection Agency (EPA). System of registries. http://iaspub.epa.gov/sor/intro$.startup.

[(EP00] Patrick Garvey (EPA). Conference report - 12th annual dama symposium. http://www.dama.org/public/pages/index.cfm?pageid=515, 2000.

[FAO] FAO. Food and agriculture organization of the united nations. http://www.fao.org/.

[fE00] European Commission Directorate-General for Environment. Guidance document for eper implementation. Technical report, European Communities, 2000.

[Fir98] Joseph M. Firestone. Dimensional modeling and e-r modeling in the data warehouse. *http://www.dkms.com/papers/dmerdw.pdf*, White Paper No. Eight, 1998.

[Giv] Kerry Given. Gis. http://www.aboutconstruction.org/GIS.php.

[Gov05] Australian Government. State of the environment australia. http://www.deh.gov.au/soe/about.html, 02 2005.

[Gro] NCSU Water Quality Group. Watershedss - a decision support system for non-point source pollution control. http://www.water.ncsu.edu/watershedss/.

[GT95] Sean Gordon and Dan Tunstall. Environmental information - the creation and distribution of environmental information. *http://www.asis.org/Bulletin/Apr-95/gordon.html*, 1995.

[HHD98] Mark W. Humphries, Michael W. Hawkins, and Michelle C. Dy. *Data Warehouse - Architecture and Implementation*. Prentice Hall PTR, 1998. ISBN 0130809020.

[HRR00] Traci Hess, Loren Paul Rees, and Terry Rakes. Using autonomous software agents to create next generation of decision support systems. 2000.

[Hyd06] BC Hydro. 2006 bc hydro annual report. http://www.bchydro.com/rx_files/info/info47209.pdf, 2006.

[IAP] IAPMEI. Dr 193 - sÉrie i-a. http://www.iapmei.pt/iapmei-leg-03.php?lei=2021.

[Ins] Rudjer Boskovic Institute. Decision trees. http://dms.irb.hr/tutorial/tut_dtrees.php.

[KAD99] Peter Kristensen, Lloyd Anderson, and Nickolai Denisov. A checklist for state of the environment reporting. Technical report, European Environment Agency (EEA), 1999.

[KC04] Ralph Kimball and Joe Caserta. *The data warehouse ETL toolkit : practical techniques for extracting, cleaning, conforming, and delivering data.* Wiley Publishing, Inc, 2004.

[Kee04] Peter Keenan. Using a gis as a dss generator. *http://dssresources.com/papers/features/keenan/keenan12172004.html*, 2004.

[Kim] Ralph Kimball. Indicators of quality. *http://www.intelligententerprise.com/000410/webhouse.jhtml?_req*

[Kim96a] Ralph Kimball. Dealing with dirty data. *http://www.dbmsmag.com/9609d14.html*, Setembro 1996.

[Kim96b] Ralph Kimball. Slowly changing dimensions. *http://www.dbmsmag.com/9604d05.html*, Abril 1996.

[Kim98] Ralph Kimball. Surrogate keys. http://www.dbmsmag.com/9805d05.html, 1998.

[Kim99] Ralph Kimball. When a slowly changing dimension speeds up. *http://www.iemagazine.com/db-area/archives/1999/990308/warehouse.jhtml? _requestid=666176*, II, 1999.

[Kim03] Ralph Kimball. The soul of the data warehouse, part 3: Handling time. *http://www.intelligententerprise.com/030422/607warehouse1_1.jhtml?_requestid=681351*, Abril 2003.

[KR02] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit*. John Wiley and Sons, Inc., 2002. ISBN 0-471-20024-7.

[Lou] GIS Lounge. What is gis? http://gislounge.com/library/introgis.shtml.

[LP04] Ana Sofia Lopes and João Moura Pires. Análise dos dados dos registos europeus - eper 2002 (relatório interno). Technical report, Instituto do Ambiente, 2004.

[LS] Lda Linde Sogás. Recuperação de voláteis. http://www.linde.pt/international/web/lg/pt/likelgpt.nsf/DocByAlias/ind--amb-cov.

[Mat06] Rosa Matias. Integração de informação geográfica em sistemas olap. Master's thesis, FCT-UNL, 2006.

[MNL] C. Di Mauro, J.P. Nordvik, and A.C. Lucia. Multi-criteria decision support system and data warehouse for designing and monitoring sustainable industrial strategies an italian case study. *http://www.iemss.org/iemss2002/proceedings/pdf/volume%20tre/241_dimauro.pdf*.

[Mor] Aran Bey Tcholakian Morales. Sistemas de apoio a decisão. http://www2.stela.ufsc.br/aran/sad/sad-aula7.htm.

[MV01] Ana Moura and Marcio Vitorino. Using mediator and data warehouse technologies for developing an environmental decision support system. *http://www.ipanema.ime.eb.br/ãnamoura/publicacoes.html*, 2001.

[Nat06] United Nations. United nations statistics division. http://unstats.un.org/unsd/default.htm, 2006.

[Obs] NASA Earth Observatory. The carbon cycle. http://earthobservatory.nasa.gov/Library/CarbonCycle/carbon_cycle4.html.

[OEC06] OECD/IEA. International energy agency. http://www.iea.org/, 2006.

[oEQT] Texas Commission on Environmental Quality (TCEQ). Decision support system for air operating permits. http://www.tceq.state.tx.us/permitting/air/nav/air_supportsys.html.

[oISS96] NATIONAL UNIVERSITY OF SINGAPORE Department of Information Systems and Computer Science. Model management for decision support. http://www.comp.nus.edu.sg/ yeogk/MM/overview/, Julho 1996.

[OMG] OMG. Unified modeling language. http://www.uml.org/.

[Org04] NISO Press National Information Standards Organization. Understanding metadata. *http://www.niso.org/standards/resources/UnderstandingMetadata.pdf*, N/A:20, 2004.

[oS99] The National Academy o Sciences. Appendix d - example prototypes. http://newton.nap.edu/html/geolibraries/app_d.html, 1999.

[Pan] Bwxt pantex. www.pantex.com/.

[PCTM03] John Poole, Dan Chang, Douglas Tolbert, and David Mellor. *Common Warehouse Metamodel (CWM) Developer's Guide*. Wiley Publishing, Inc., 2003.

[pm] Ecobase project members. The ecobase project: Database and web technologies for environmental information systems. *www.sigmod.org/record/issues/0109/a2-pvur.pdf*.

[POR] SOLVAY PORTUGAL. Desenvolvimento sustentável. http://www.solvay.pt/sustainabledevelopment/0,,1210-5-0,00.htm.

[Pow03] D. J. Power. A brief history of decision support systems. dssresources.com, world wide web. *http://dssresources.com/history/dsshistory.html*, version 2.8, Maio 2003.

[Proa] United Nations Environment Programme. Barcelona convention. http://www.unep.ch/regionalseas/regions/med/t-barcel.htm.

[Prob] United Nations Environment Programme. United nations environment programme - unep. http://www.unep.org/.

[RCVT06] D. Rapti-Caputo, C. Vaccaro, and L. Teruggi. Rio quequén grande basin (argentina): impact of geological, environmental and human activities on water quality. *Geophysical Research Abstracts*, Vol. 8, 2006.

[REA05] RelatÓrio do estado do ambiente 2003. Technical report, Instituto do Ambiente, 2005.

[Rel01] Relatório final de projeto institucional cnpq/inria. vega.cnpq.br/pub/protem/workshop2001/inria/relatorios/Ecobase.rtf, 2001.

[S.A] Solvay S.A. Relatórios ambiente em portugal. http://www.solvay.pt/solvayinportugal/plantofpovoa/relatorioambienteportugal/0,,2262-5-0,00.htm.

[SKL] Jeff N. Stovall, David Korns, and Clarence Lamb. Environmental data warehouse integrates analytical data, gis, and the web. *http://gis.esri.com/library/userconf/proc03/p0912.pdf*.

[Sni01] Ronald Snijder. Metadata standards and information analysis. *http://www.geocities.com/ronaldsnijder/*, 2001.

[Son04] Marc Songini. Etl. *http://www.computerworld.com/databasetopics/ businessintelligence/datawarehouse/story/ 0,10801,89534,00.html*, Fevereiro 2004.

[SS] Earth Observatory NASA/GSFC Security and Privacy Statement. The carbon cycle. http://earthobservatory.nasa.gov/Library/CarbonCycle/carbon-cycle4.html.

[Sys] Inmon Data Systems. Corporate information factory. http://www.inmoncif.com/home/.

[Szn03] Ronald J Sznaider. Solving complex problems with gis and advanced meteorological data. *http://ams.confex.com/ams/annual2003/techprogram/paper_54964.htm*, 2003.

[tgtgis] the guide to geographic information systems. What is gis? http://www.gis.com/whatisgis/index.html. gis.com - the guide to geographic information systems.

[UA97] UNEP/GRID-Arendal. State of environment norway 1997. http://www.grida.no/prog/norway/soeno97/aboutsoe/aboutsoe.htm, 05 1997.

[USG] USGS. Geographic information systems. http://erg.usgs.gov/isb/pubs/gis_poster/.

[UW] UNEP-WCMC. United nations environment programme - world conservation monitoring centre. http://www.unep-wcmc.org/.

[Var02] Ganesh Variar. The origin of data. *http://www.intelligententerprise.com/020201/503feat3_1.jhtml*, Fevereiro 2002.

[Wika] a enciclopédia livre. Wikipédia. Erp. http://pt.wikipedia.org/wiki/ERP.

[Wikb] Wikipedia. Eutrofização - wikipedia. http://pt.wikipedia.org/wiki/Eutrofica%C3%A7%C3

[Wikc] the free encyclopedia Wikipedia. Executive information system. http://en.wikipedia.org/wiki/Executive-Information-Systems.

[Wikd] the free encyclopedia Wikipedia. Xml metadata interchange. http://en.wikipedia.org/wiki/XML-Metadata-Interchange.